

Coping with a budget reversal

Last week's announcement by the German government of budget cuts in research are unwelcome but hardly surprising. The leaders of the research community need to focus on long-term restructuring to make the most of declining funds.

The impression of decline is palpable in autumnal Germany. The inability of the Social Democrat–Green government, re-elected in September, to combat the crisis—a vicious circle of an ageing society, high unemployment, exploding costs of the social-security system, high taxes and weak consumer demand—dominates the newspaper headlines and political talk shows. Is the third-largest economy in the world really the weakest in Europe?

Widespread expressions of panic and the readiness, or glee, with which the media joined in the chorus of Cassandras are disturbing aspects of the crisis, which some commentators have compared to the decline of the Weimar Republic in the 1920s that helped fascism to gain power. But Germany's current slackness has little to do with a failure of the political system, and a lot to do with mass psychology and a lack of citizens' trust in the country's economic future.

Public spending is too high. Ironically, Germany, once a paragon of economic growth and stability, is likely to be sanctioned next year by the European Union (EU) for its failure to meet the Maastricht criteria for stability. Under these circumstances it is less unexpected than some science administrators would like to admit that the government has backed down from a pre-election promise to increase the budgets of Germany's main scientific organizations (see page 452).

But the decision comes at a precarious time for the German research system, which is in the midst of a difficult reform process. Two of its strongest pillars, the Deutsche Forschungsgemeinschaft (DFG)—the principal funding agency—and the elite research centres of the Max Planck Society (MPS), have been thoroughly evaluated in recent years. Both have since set up new initiatives aimed at fostering collaboration between researchers in and outside universities, and at increasing training opportunities for the best young scientists.

It would be a pity if these promising new schemes, such as the International Max Planck Research Schools and planned investment in high-profile research in eastern Germany, were to be halted.

Cutting the most recent, and least protected, achievements would be the path of least resistance. But there are better ways to respond to the current situation.

There is no shortage of evaluation of institutions and diagnosis of weaknesses, but the promised evaluation of the system as a whole has not been achieved. The autonomy of universities, in terms of employment and entrepreneurial freedom to pursue new opportunities in research and education, is under-developed. And while implementing change at universities is a challenge, the main pillars of non-university research—the MPS, the Hermann von Helmholtz Association of National Research Centres, the Fraunhofer Society for applied research, and the Leibniz Association—remain too isolated. These opportunities would, in the long term, allow Germany's research base to save far more than the €100 million (US\$100 million) or so that the science organizations need to save next year.

European statistics published last week show that in the 1990s Germany managed to maintain and strengthen its scientific productivity, while creating an entire research base in the east. But the MPS, the DFG and other German research organizations must strengthen their mechanisms of quality control to be able to allocate resources even more effectively.

The age structure of the MPS provides an opportunity for the society to smoothly renew itself. One-third of its scientific directors will retire in the next few years—a period it must use to carefully rethink its priorities. A smaller and healthier MPS is preferable to an over-staffed apparatus with little money left for investment.

In the long term, the image abroad of Germany will not depend on whether or not its research receives 3% more in 2003. What will make a difference, however, is whether or not the country manages to tackle the structural problems that it has allowed to accumulate over the past quarter-century. If Germany's universities and research organizations lead the way, so much the better. ■

Promoting animal research

Researchers need to be more active in explaining the value and necessity of their work.

The odds are that there will ultimately be a favourable outcome to a planning inquiry currently under way in Cambridge, UK, examining the case for a new centre for primate neuroscience at the university. It is being held under the auspices of the local authority, which previously decided against the centre for the unusual reason that the probable public protests would be too problematic. The outcome will be referred (probably early next year) to the British deputy prime minister, John Prescott, for a final decision, and the government has rightly declared the centre to be a national 'need'—thus, incidentally, justifying construction on a site of protected 'green field' status.

The arguments being deployed in the inquiry are not about science. However, opponents are disputing the 'need' on the quasi-scientific and fallacious grounds that animal research is both misleading and unnecessary, given available alternatives.

What is striking about the debates is the near-invisibility of the

scientific community, apart from the stalwart but predictable Research Defence Society. This absence of active researchers does nothing but help the opponents of essential research that has in the past attracted broad public acceptance. Certainly, some animal researchers have been treated viciously in other contexts, but objectors here have restricted themselves to presenting the arguments.

The evidence of opinion polls and the lack of public hostility to some people who have recently supported animal research suggests that a well-planned campaign of information and public representation can keep the worst excesses at bay. Relying on individual researchers to stand prominently in isolation is a recipe for scientific and democratic failure. It is feasible for scientists to band together to campaign on a single issue. Whatever happens at Cambridge, collective public representation in Britain and elsewhere of the animal researchers' case needs to be developed in a professional manner. ■





Net loss
US plans for deep-sea
aquaculture draw
sharp criticism
p451



Frozen assets
East German
research falls victim
to cash shortage
p452



Home grown
Transgenic trials
don't crop up in
science review
p453



Rumbling spire
Submerged volcano
prepares for trip
to the surface
p455

Surgeons struggle with ethical nightmare of face transplants

Natasha McDowell, London

Transplant experts are this week coming to terms with claims by a top London surgeon that face transplants could soon move from the realm of Hollywood fantasy into actual clinical practice.

The technology needed to perform a face transplant could be ready within six months, Peter Butler of the Royal Free Hospital told the winter meeting of the British Association of Plastic Surgeons in London on 27 November.

"The science is a small part. The question is not can we do it, but should we do it," Butler said. He added that hand transplants have already provided proof of principle that composite tissues of skin, blood and bone can survive transplantation with the help of modern immunosuppressive drugs. Such drugs suppress the recipient's immune system and so prevent the grafted tissue from being rejected by the patient's body.

For many, the claim may bring to mind the face- and identity-swapping escapades of John Travolta and Nicholas Cage in the 1997 movie *Face/Off*. But in real life, the transplant issue will arise for people who suffer from disfigurements that are so severe as to make the drastic surgery a viable consideration.

Some of these people find Butler's statements unhelpful. "It is a speculative idea, but has gained huge media attention that may have been damaging to individuals trying to adjust to their disfigurement," says burns victim James Partridge, founder and chief executive of the London-based charity Changing Faces, which helps those with facial disfigurement. "It should have been opened as a private debate between psychologists and surgeons before being made public." Partridge says that he has not met anyone who is interested in undergoing a face transplant. "Even for those like myself with severe disfigurement, the face carries a lot of identity, the sense of self," he says.

But Christine Piff, founder of the UK support group Let's Face It, says the technology could benefit those whose disfigurement is so severe that they become withdrawn, or those who cannot eat, drink or speak. She herself lost half of her face to cancer 25 years ago, and



Hand up: hand transplants have already shown that tissue composites can survive transplantation.

adds that although she would probably have refused a transplant at the time, when she reflects on the advantages it might have offered, she is now not so sure.

Butler says that although the ultimate aim is to transplant bone and muscle, his team's initial approach will be to transplant only skin, subcutaneous tissue and blood vessels. This would leave patients with their own recognizable features, but with potentially more flexibility than is currently available with a graft of a patient's own non-facial skin.

Critics say that practical use of the approach is further off than Butler suggests. "Technically, face transplants could be done this afternoon," says John Clarke, a plastic surgeon and director of the burns unit at Chelsea and Westminster Hospital in London. "The big problem will be finding donors, and the psychological difficulties could be dreadful."

Any patient who received the transplant would have to take immunosuppressive drugs, with their increased risk of cancer and infection, for the rest of his or her life. For this reason, the treatment would only be ethically acceptable either for patients who are so

disfigured that they become suicidal, or for those who already have life-threatening facial cancer. But immunosuppressive drugs do not always prevent rejection, and could hinder the patient's ability to combat the cancer, says Jim Frame, a plastic reconstructive surgeon who has spent 20 years working on the possibility of face transplants. He now thinks that more promising approaches will come from the use of man-made materials to restore patients' facial features.

However, Butler's hope is that immune rejection may eventually be overcome by an approach being developed at Massachusetts General Hospital in Boston. This involves first grafting bone marrow from a donor while treating the patient with immunosuppressives. Work on animals suggests this can induce immune tolerance in the recipient, allowing a further transplant from the same donor without risk of rejection.

But this approach is currently aimed at those with bone-marrow cancer who are at risk of accompanying organ failure, says Steve Schey, director of the stem-cell transplant unit at Guy's Hospital, London. He says the approach is interesting, but speculative. ■

J. LAIR, JEWISH HOSP.; KLEINERT, KUTZ & ASS.; HAND CARE CENTER; UNIV. LOUISVILLE/NEWSMAKERS

Prion data suggest BSE link to sporadic CJD

Declan Butler

Predicting the number of cases of Creutzfeldt–Jakob disease (CJD) in people as a result of transmission of bovine spongiform encephalopathy (BSE) has just got more difficult.

Whereas it was thought that BSE only caused a new form of the disease called variant CJD (vCJD), a study in mice from a team led by John Collinge at University College London suggests that it may also cause a disease indistinguishable from the commonest form of classical, or 'sporadic', CJD (E. A. Asante *et al.* *EMBO J.* **21**, 6358–6366; 2002). If the group's mouse model is relevant to the

human disease, the results also suggest that the true extent of infection may be difficult to assess because of the large number of asymptomatic carriers.

The latest work uses mice engineered to carry the human gene for a cell-membrane protein called PrP. Prion diseases occur when PrP is converted to the abnormal 'prion' form, PrP^{Sc}. Collinge has developed a test, based on a standard western blot for analysing proteins, to study PrP^{Sc} extracted from the brain. This previously showed disease caused by BSE or vCJD to give a characteristic molecular signature that is distinct from sporadic CJD (J. Collinge, K. C. L. Sidle, J. Meads, J. Ironside and A. F. Hill *Nature* **383**, 685–690; 1996).

In their latest experiments, Collinge and his team injected material from the brains of cows with BSE or people with vCJD directly into the brains of two strains of mice with a human PrP genotype known as 129MM. Almost 40% of the British population has this genotype. In one mouse strain, those that became infected showed the usual BSE pattern in the western blot. But in the other, Collinge's team tested 11 mice infected with BSE material using the western blot. Ten of them showed a pattern consistent with sporadic CJD.

The number of cases of sporadic CJD have been rising in Britain since the 1970s, and this had been attributed to better monitoring for the condition. But in July, researchers led by Adriano Aguzzi of the

University Hospital Zurich reported a sudden increase in sporadic CJD figures in Switzerland in 2001, and suggested that infection with BSE might be to blame (see *Nature* **418**, 266; 2002). Collinge's new data provide worrying molecular evidence that BSE might be to blame for the rise in sporadic CJD.

In previous experiments, Collinge had injected BSE and vCJD material into mice with another human PrP genotype, known as 129VV. The new data are thought to be more relevant to the transmission of BSE because all of the known human victims of vCJD have the 129MM genotype. Another worrying finding is that the 129MM mice seem to be more susceptible to developing a subclinical infection, with no obvious symptoms.

If a large pool of the British population is carrying a subclinical BSE infection, this would have serious consequences for the potential transmission of the disease, for instance through contaminated surgical instruments. And although laboratory mice are short-lived, infected humans might go on to develop the disease later in life.

The UK Department of Health, which has been briefed by Collinge on his findings, says that it will ask its Spongiform Encephalopathy Advisory Committee to consider the results closely at its next meeting in February. Collinge says that an urgent nationwide screening of tonsil material is needed to get a better estimate of the level of infection in the population. ■



Could transmission of BSE to humans be responsible for more than one type of CJD?

Europe urged to provide boost for bioterror research

Alison Abbott, Berlin

Bioterrorism may be the word on everybody's lips at the moment, but Europe needs to put its money where its mouth is, researchers at the continent's first scientific meeting on the topic were told last week.

A lack of investment in bioweapons research could leave Europe vulnerable to attack, speakers at the meeting asserted.

Under the European Union's Sixth Framework Programme for research funding, "virtually no money is specifically earmarked for such research", said Stefan Kaufmann, a director of the Max Planck Institute for Infection Biology in Berlin. By contrast, the US National Institutes of Health has earmarked more than \$1.5 billion for a very broad programme, he added.

"If there were a bioterrorist incident in Europe there would be uproar," Kaufmann warned the meeting, "so

politicians need to take the issue of research seriously — and now."

The meeting, a symposium on bioterror and bioweapons held in Berlin on 30 November–1 December, came as Britain revealed that it plans to stockpile about 60 million doses of smallpox vaccine in preparation for a bioterrorist attack.

The sixth Framework, which will issue its first call for applications later this month, supports research into infectious diseases, but focuses on the three conditions linked to poverty — tuberculosis, malaria and AIDS — rather than those with weapons potential, such as anthrax and Ebola. Scientists can apply for funds to study the latter, but only under narrow programmes such as 'rational development of antiviral drugs' or 'combating resistance to antibiotics'.

But researchers at the meeting questioned whether the huge infusion of funds into US bioterrorism research will be



Stefan Kaufmann: Europe must be prepared.

used effectively, and said that it should not necessarily be matched in Europe. Critics have also expressed fears that the expansion of such research in the United States will heighten the risk of bioterrorism, by spreading materials and specialist knowledge.

Milton Leitenberg, an arms-control expert from the University of Maryland, College Park, questioned the

fundamental philosophy of large programmes specifically targeting bioterrorism. "There would be a greater spin-off for military purposes by putting money into diseases that are the biggest public-health threat, such as HIV and tuberculosis," he said. ■

US pushes fish farming into deep water

Rex Dalton, San Diego

The US government is set for a confrontation with environmentalists as it investigates the viability of deep-sea fish farms in federally controlled waters.

The farms could be built more than 5.5 kilometres off the US coast — where the regulatory reach of individual states generally ends — and would keep fish in large, diamond-shaped cages about 30 metres beneath the ocean surface.

If the plan goes ahead it is likely to spark fierce opposition not just from environmental groups, but also from state officials, some of whom are already worried about the impact of shallow-water aquaculture on natural species.

The suggested sites for the farms lie in the US Exclusive Economic Zone, which extends from the edge of US territorial waters to 370 kilometres offshore. Aquaculture companies are already engaged in a “gold rush” for access to these offshore sites, says John Volpe, an ecologist at the University of Alberta in Edmonton, Canada, who is familiar with the plans. “It’s like the Wild West,” he says.



Environmentalists are worried by plans to extend aquaculture to beyond US territorial waters.

Advocates of deep-water aquaculture claim that results from preliminary research indicate that the concept is not environmentally damaging, and might reduce the ecological problems seen in near-shore fish farms. This is because the deep-water farms will be more dispersed, and further from the spawning grounds for some wild fish.

But the limited number of scientists and

government officials aware of the plans are expressing concern. They fear that pollution from penned fish, spread of disease or parasites to wild populations, or habitat harm from inadvertent release of caged species all could harm wild fish.

Volpe says he is worried that those involved in the planning process will repeat the mistakes of near-shore salmon farming in British Columbia, where escaped and diseased fish are blamed for ecological damage.

To counter this, a number of federally funded pilot projects studying deep-water aquaculture are planned or are under way. One, examining cobia, is some 35 km off the coast of Mississippi in the Gulf of Mexico.

The National Oceanic and Atmospheric Administration (NOAA) is spending \$600,000 on a study at the Center for the Study of Marine Policy at the University of Delaware to develop a policy and regulatory framework for fish farming between 5.5 and 370 km offshore. The three-year study will be completed in the spring, and proponents hope it will convince Congress to pass legislation allowing commercial aquaculture in this region.

And NOAA's National Marine Fisheries Service has already held workshops to develop a voluntary code of conduct for future commercial operations in the offshore belt. There also are plans to put this voluntary code into federal law, officials say.

“We are trying to fund the science for an environmental understanding of the concept,” says James McVey, a marine ecologist at NOAA's Sea Grant College programme in Silver Spring, Maryland. “We hope to foster new technologies to bring to the world and the United States.”

But in Alaska, where aquaculture is not permitted in state waters, scientists and government officials are fighting the plans. “NOAA should not promote aquaculture in federal waters adjacent to any state that has laws banning the practice,” wrote Governor Tony Knowles in a letter to the National Marine Fisheries Service.

Pathogen-tracking questioned

Erika Check, Washington

The US government's main programme to keep track of dangerous pathogens used in laboratories is not being run effectively, according to a Congressional investigation.

The Centers for Disease Control and Prevention (CDC), based in Atlanta, Georgia, administers the Laboratory Registration/Select Agent Transfer Program. Under the scheme, the CDC maintains a list of deadly organisms and toxins, such as anthrax and botulinum toxin, and tracks their transfer between the hundreds of labs, at universities and elsewhere, that use them in research.

But according to the General Accounting Office (GAO), which began a review of the programme in November 2001 after the still-unsolved anthrax attacks, the CDC must do a better job of keeping tabs on these labs.

The GAO's findings were released just as the CDC prepares to revamp and increase the scope of the select agents programme.

“We found significant management weaknesses in CDC's facility registration and transfer monitoring processes that impede effective program oversight,” the GAO wrote in a letter on 22 November to Tommy Thompson, head of the Department of Health and Human Services, of which the CDC is a part.

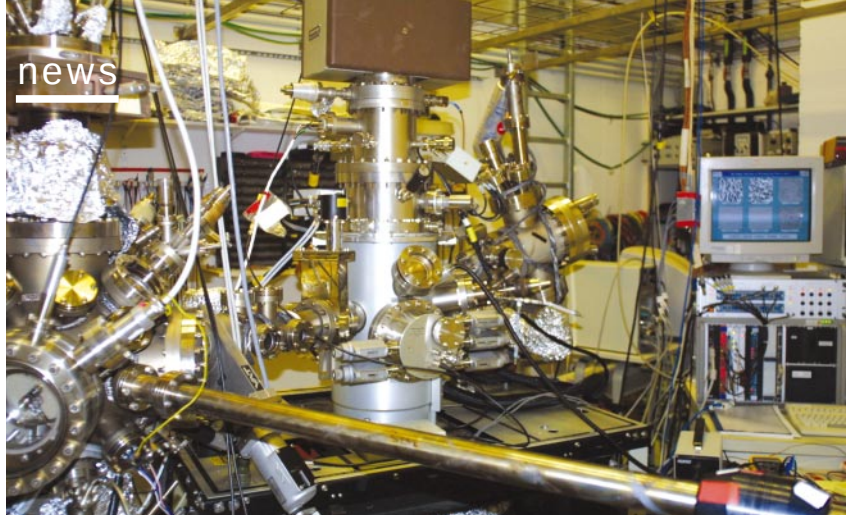
Observers say that the issues raised in the report may be addressed soon, when

new regulations take effect. In June, President George Bush signed a law that, among other things, requires the CDC to keep track of all labs that use select agents — not just those transferring the materials. On 9 December, the government will publish for public comment the regulations required by this plan, says Stephen Ostroff, deputy director of the CDC's National Center for Infectious Diseases.

The select agents programme has never been very popular: scientists resent the paperwork that it creates, and the CDC doesn't like regulating labs with which it would rather collaborate.

The CDC has also encountered delays and serious problems in administering the programme, the GAO says. Early versions of the legislation that created the new Department of Homeland Security would have wrested the entire programme from the CDC's control. But the final bill left the programme's administration unaltered. Ostroff says that the CDC has already begun to respond to the GAO's concerns by hiring more contract workers to carry out inspections, and by allocating funds to a new database.

“Everybody views this as an essential component of national security, and we're committed to running the programme as well as we can,” Ostroff says.



Hard times: the Max Planck Institute of Microstructure Physics must sell its surface-analysis kit.

Funding freeze leaves eastern Germany out in the cold

Quirin Schiermeier, Munich

Ever since German reunification in 1990, strenuous efforts have been made to beef up science in the former East Germany. The Max Planck Society (MPS) and others have injected money and talent to create institutes there of international standing. But an announcement on 30 November that funding for German science will be frozen next year has raised fears that the build-up will now falter.

The funding freeze, announced by science minister Edelgard Bulmahn at a meeting with the heads of Germany's main research organizations, is a heavy blow for the MPS, which runs 80 basic research institutes. Since reunification, the society has founded 20 institutes in eastern Germany. Now it faces a €50-million (US\$50-million) reduction in its previously projected 2003 budget of €1.2 billion, forcing it to revise its scientific goals.

The Max Planck Institute for Biogeochemistry in Jena, south of Berlin, for example, recently reached an agreement with the Russian government over the construction of a 250-metre tower for measuring carbon dioxide fluxes over Siberian forests. But this new project may have to be dropped.

"The tower would offer unique scientific opportunities at one of the hotspots of global warming," says Ernst-Detlef Schulze, the institute's scientific director. "It would hit us hard if it were stopped."

The 'Aufbau Ost' — the campaign to build up eastern Germany since 1990 — has enabled the MPS to enter new scientific territory. Many of the generously equipped institutes are becoming internationally competitive in fields such as cell biology, genetics, evolutionary anthropology, primate research and plant biology.

But that financial largesse is now disappearing, says Jürgen Kirschner, scientific

director of the Max Planck Institute of Microstructure Physics in Halle.

The institute's facility for magnetic surface analysis, for example, which cost €500,000 to build, is to be sold to a US laboratory because its operating costs are higher than expected and its maintenance support is falling.

"Running and repair costs are eating us up," says Kirschner, who has little hope that the MPS will now invest in a hoped-for new €750,000 atomic-force microscope for nanostructure analysis.

Cutting investment is one means of coping with the budgetary constraints, says Peter Gruss, the MPS president. But in the longer term, the society will only be able to maintain scientific quality by reducing the number of its research departments, he says. "We will need to close at least 20 departments, maybe even some institutes," he says. "This means a 10% reduction in our research potential."

Ten of the new institutes in the east are still without directors, partly because of the difficulties of luring scientific heavyweights to eastern Germany. Some of these positions may remain vacant, says Gruss.

Across Germany, a third of the society's 240 scientific directors will reach retirement age in the next six years. Gruss says that only 60 of the resulting 80 vacancies will be filled.

The cutbacks in the 2003 budget plans for Germany's research organizations were announced as the country faces its worst economic crisis for decades. Even so, the government's decision to renege on a pledge made before September's election of a 3% increase in research budgets next year surprised some scientists and research administrators.

"This is a drastic turn in research policy," says Ernst-Ludwig Winnacker, president of the DFG, Germany's main university research funding agency, which will have to save millions of euros next year (see story, right). ■

Postdoc positions axed as economic crisis takes its toll

Quirin Schiermeier, Munich

As many as 2,000 young German scientists are set to get a nasty shock in the mail as laboratories cancel planned positions in response to a budget freeze.

After last week's announcement that research budgets will be frozen at this year's level (see left), the DFG, Germany's main university research funding agency, will have €43 million (US\$43 million) less to spend than it had anticipated.

The agency supports about 27,000 graduate students and postdoctoral fellows in all disciplines. It says that this number is set to fall by about 10%.

"The success rate of applicants will decrease," says Ernst-Ludwig Winnacker, the DFG's president. Most disappointed will be those young researchers whose planned thesis or postdoctoral work in a collaborative project has been approved and must now be cancelled, he adds.

The Max Planck Society is not doing any better. The planned extension of International Max Planck Research Schools — an initiative aimed at improving training opportunities for PhD students — is likely to be delayed, and plans to set up a Max Planck group for stem-cell research at the University of Ulm may also be suspended.

Additionally, the continuation of the European Neuroscience Institute in Göttingen — four independent junior neurobiology groups jointly funded by the Georg August University of Göttingen and the MPS — may be at risk.

Stopping the project now would be disastrous, says Erwin Neher, who runs the Max Planck Institute for Biophysical Chemistry in Göttingen. "What's the use of complaining that Germany's brightest talent leaves the country if you cut, of all the institutes, the most attractive and competitive ones?" ■



Field trials excluded from UK crop appraisal

David Adam, London

Britain's field-scale trials of genetically modified crops cost millions to set up and are the most extensive studies of their kind in the world. So plant scientists were surprised to learn last week that their results will not be included in a comprehensive, government-funded science review of the safety of such crops.

The review is one of three initiatives the government has scheduled for the run-up to its decision on whether to allow the commercial planting of transgenic crops. The other two are a study of the economic benefits and a national public debate (see *Nature* **419**, 327; 2002). The government wants the results of all three initiatives by June next year, a month before the first results from the field-scale trials are due to be published.

The trials will assess the effect on wildlife of large-scale cultivation of the crops (see *Nature* **412**, 760–763; 2001). “These field releases are the only large-scale risk assessment that’s been carried out in Britain and it’s odd that we’re not going to discuss them at all,” says Carlo Leifert, a nutritionist who directs the Tesco Centre for Organic Agriculture in Stocksfield, Northumberland, and is a member of the panel that will conduct the science review. “It’s one of the things I intend to raise at our first meeting.”

Government officials say that the trials were not being included in the review because they are only addressing a narrow question about one type of crop and bio-



Big study, small result: government officials say that field-scale trials address only narrow questions.

diversity. “It would be absurd if we focused on one trial in the United Kingdom,” says David King, the government’s chief scientific adviser. “I don’t think many countries know about our little experiments.”

But such language contrasts sharply with previous government statements. In March 2000, environment minister Michael Meacher said: “There will be no commercial growing until we are satisfied there will be no unacceptable effects on the environment. That is why we have led the world by setting up this research programme, which will give us the answers to these important questions.”

The government says that it will still consider the results of the trials before making a decision on whether to allow commercial use of transgenic crops, but many still do not understand why the results are to be omitted from the science review.

Researchers advising the government could only say that they have come under pressure to ensure that the review and debate are finished before the trials’ results come out. “They were very clear that they wanted it over and done with by June,” says one member of the steering group that will coordinate the public debate. ■

Royal Institution’s director blasts scientific sexism

David Adam, London

Pity Danielle Ohayon, otherwise known as Miss Jamaica, a contestant in this year’s Miss World competition. Already forced to relocate to London following violent protests against the beauty pageant in Nigeria, Ohayon received further bad news last week — she faces institutional sexism if she follows her desired career in marine biology.

Likewise, if Mai Phuong Pham Thi — Miss Vietnam — lives out her dream and becomes a physicist, she may find it difficult to return to work if she wants to take a break to have children. And Chinenye Ochuba — Miss Nigeria — is likely to strike the long-standing and notorious glass ceiling should she achieve her ambitions in computer science.

The gloomy prognosis of young women’s prospects in science arrived in a report from Susan Greenfield, director of the London-based Royal Institution of Great



Miss vocation: aspiring biologist Danielle Ohayon.

Britain. Commissioned by the British government and published on 28 November, the report calls for state funding for fellowship schemes to retrain women who have taken breaks in their career to start a family.

It also calls for funds for family-friendly practices such as part-time working and job-share programmes, which it says would help women to rise through the research ranks.

“Thankfully we have gone beyond bottom pinching, but in some ways the latest form of discrimination is worse,” Greenfield says. “It’s hidden, institutional sexism, clearly reflected in the awful statistics.” Just 8% of women academics

in Britain’s older universities are professors, for example, and more than a third are on the lowest lecturer pay scale. The report points out that there are similar problems in other countries, echoing accounts that sexism persists in the United States, Japan and elsewhere.

The low participation of women in British science is not just a problem for them, the report says, but also for the country, business and society as a whole.

It adds that many of the efforts and initiatives to encourage greater participation are fragmented — and that many women in science are unaware that they even exist. To address this problem, the report suggests that the government should set up a ‘working science centre’ to draw these programmes and schemes together. The government says that it intends to consider the recommendations and will publish a full response shortly. ■

Germany sets new criteria for reinstated gene-therapy trials

Munich Gene-therapy trials using retroviral vectors are to continue in Germany, ending a moratorium introduced in June.

In October, trials were suspended in several countries after news emerged that a French child involved in a trial of gene therapy to treat severe combined immune deficiency had developed a form of leukaemia, which experts believe was probably triggered by the retroviral gene vector. German authorities, however, had already suspended a series of trials following similar results in an animal experiment (see *Nature* **419**, 545–546; 2002).

The German Chamber of Physicians, together with the Paul Ehrlich Institute in Langen, which authorizes gene-therapy trials, now says that trials involving retroviral vectors can continue, provided that patients are made aware of the risks, and that they stand to benefit from the treatment. Three of the suspended trials — one for the immune disorder chronic granulomatous disease, and two in graft-versus-host disease — have been given the green light. Two new protocols, for trials in HIV infection and rheumatoid arthritis, are now also being considered.

The organizers of the other 11 interrupted

German trials have not submitted amended or new protocols, and the trials will not continue. Many of these would have been unable to fulfil the new criterion of patient benefit, says Klaus Cichutek, vice-president of the Paul Ehrlich Institute.

Researchers alarmed by 'clone pregnancy' hype

Rome Severino Antinori, the maverick Italian fertility specialist, was again courting controversy last week, claiming that three of his women patients are carrying cloned babies — one, he says, is 33 weeks pregnant.

Mainstream scientists are sceptical of the announcement, pointing out that it is at odds with reports in April that said Antinori had



Severino Antinori's claim to have three patients carrying cloned babies has met with scepticism.

claimed at a meeting in the United Arab Emirates that one patient was already two months pregnant (see *Nature* **416**, 570; 2002). Antinori's previous unverified claims include a 1999 announcement that he had produced test-tube babies using sperm from infertile men matured in the testes of rodents.

Nevertheless, researchers fear that the continuing negative publicity could influence legislatures considering cloning, perhaps causing them to ban researchers from using the technique to produce human embryonic stem cells for studies of regenerative medicine.

Holed reactor causes leak in fusion finances

Tokyo Call it an expensive puncture: Kyushu University's Advanced Fusion Research Center in Japan is having to shell out ¥30 million (US\$240,000) to repair a 0.15-mm hole in its 1.7-m-diameter tokamak reactor.

Inside the fusion reactor, hydrogen is held as a plasma, confined by a magnetic field, and heated to temperatures reaching 50 million °C. The hole was found in July, but the news only emerged earlier this week. It is being blamed on an accident in which the heated plasma was mistakenly allowed to touch the vessel walls. "It happens from time to time, but we didn't know that it would create this kind of damage," says

AFP

Naoaki Yoshida, the centre's director.

The damage will be repaired using a remote-controlled welding device, and the reactor should be as good as new by early next year, says Yoshida.

French researchers revolt over 'unfair' bonus scheme

Paris *Egalité, fraternité* — and hard cash. These factors are at the centre of a row at INSERM, France's biomedical research agency. A petition launched late last month against a new bonus scheme for the agency's 2,300 scientists has already attracted some 500 signatures.

The five-year 'interface' contracts, launched by the agency's director-general Christian Bréchet, are intended to increase mobility between the agency's labs and partners — including universities, hospitals and industry (see *Nature* 418, 908; 2002). Staff selected to take part would earn an extra □ 1,500 (US\$1,500) a month.

The petition argues that the resulting "enormous" differences in pay among similarly qualified staff would create strife. The average starting pay for INSERM researchers is barely □ 2,000 a month, the signatories point out.

This week sees the deadline on a call for proposals for 100 of the grants, and Bréchet

hopes to extend this to 800 by 2006. He says the grants will boost careers, while helping to translate basic research into clinical advances.

No easy solution to chemists' job worries

Washington Job prospects for chemistry and chemical-engineering graduates are at their worst for nearly a decade, a survey of the US market concludes. Demand is down and unemployment among chemists hit a 30-year high of 3.3% last year, says the survey for *Chemical and Engineering News (C&EN)*, published by the American Chemical Society.

Even pharmaceutical companies, traditionally a major employer of new chemistry graduates, are waiting for the economy to pick up before hiring people. The survey asked society members working or looking for jobs in the United States about their salary and status. "Job prospects for 2003 will look unhappily similar to other 'down' years, most recently the mid-1990s," says Madeleine Jacobs, *C&EN*'s editor-in-chief.

Dispute set to peak if volcano gets its head above water

Rome Earth scientists and diplomats alike have reason to watch the waters off Sicily's coast, where an island may be about to appear.



Blazing row: when Graham Bank last appeared in 1831, four nations laid territorial claim to it.

Seismic activity in the region, including the eruption of Mount Etna, could force the peak of a submerged volcano above the waves.

When the island last popped up, for six months in 1831, Britain, France, Sicily and Spain all claimed it. To scientists it is known as Graham Bank (its British name is Graham Island) but to Italians it is Ferdinandea, after the Sicilian king Ferdinand II. Italian divers reinforced their nation's claim by placing a plaque on the submerged rock in March 2001.

Francesco Italiano, a senior scientist at the National Institute of Geophysics and Volcanology in Palermo, says bad weather is hampering studies of whether the rock, which is under about 8 metres of water some 30 km off Sicily's southern coast, could reappear.

The real deal

The human genome fired the public's imagination. But for many geneticists, the genome of their main experimental mammal — the mouse — is even more exciting. *Nature's* reporters sample the buzz in three leading laboratories.

On the double

Nancy Jenkins and Neal Copeland are very much a team. Married for 22 years, they oversee the National Cancer Institute's Mouse Cancer Genetics Program in Frederick, Maryland. They have been co-authors on every paper they have published over the past two decades, many introducing new mouse models for studying cancer and other important human diseases.

Copeland and Jenkins also have a habit of finishing one another's sentences that helps to convey a sense of breathless excitement about the mouse genome. Their interest in mouse genetics began in 1980 when, as young molecular biologists, they joined the world's pre-eminent mouse lab, the Jackson Laboratory in Bar Harbor, Maine. "We decided to go there so that we could fuse molecular biology with mouse genetics to try to answer questions that would not be possible with either technology alone..." Copeland begins, "...and it turned out to be a moment that will never be recapitulated," finishes Jenkins.

While Jenkins and Copeland look back fondly on those early days, the mouse genome sequence (see page 520) is accelerating their research in ways that make their past achievements seem pedestrian. Back in the 1980s, if Jenkins and Copeland were interested in investigating a spontaneous mutation presented at the Jackson Lab's weekly 'Mutant Mouse Circus', it was a laborious process. Identifying the gene involved meant crossing about 1,000 mice to map it to a stretch of chromosome bearing about 20 candidate genes. From there, a postdoc would have to sequence all of them in both normal and mutant mice to find out which was mutated.



Labour saviour: the mouse genome sequence will allow Neal Copeland and Nancy Jenkins to spend more of their time on functional analysis.

"It used to be one postdoc project per mutation..." says Jenkins, "...and it was like looking for a needle in a haystack," adds Copeland. But since the mouse genome sequence became available in May (at www.ensembl.org/Mus_musculus), researchers can simply go to the database after the initial breeding experiments and look up all the genes in the relevant chromosomal region. By knowing from their sequences what types of proteins most of them encode, they can choose one or two that look most promising to search for the mutation.

"It took us 15 years to get 10 possible cancer genes before we had the sequence," says Copeland. "And it took us a few months to get 130 genes once we had the sequence." What's more, Jenkins points out, going back and forth between the mouse and human genomes will help to target related human genes that could be candidates for drug development.

The pair's eyes light up at the mention of the companion to the draft mouse genome: the set of full-length mouse complementary DNA (cDNA) sequences, representing an inventory of the sequences of expressed genes, produced by a team at the Institute of Physical and Chemical Research (RIKEN) Genomic Sciences Center in Yokohama, Japan (see page 563). This library of some 61,000 cDNAs gives researchers a functional copy of almost every mouse gene that can be easily manipulated in the lab — making it much easier to create gene-knockout mice, for example.

It will also erase time spent on deciphering gene sequence and structure. "We are skipping to an era that frees us from all the mechanical stuff that took up most of a mouse geneticist's

time so that we can go straight to the functional part..." says Copeland, "...and the functional part is truly the fun part!" adds Jenkins. ■

Kendall Powell

♦ <http://web.ncicfcrf.gov/research/mcgp/default.asp>

Talkin' about regeneration

Newts do it, so why don't we? That's the question that dominates the thoughts of Nadia Rosenthal, head of the European Molecular Biology Laboratory (EMBL) Mouse Biology Programme at Monterotondo, near Rome.

A newt can regrow a severed leg in its entirety, whereas mammals can only slowly patch up small-scale damage, with the production of low-quality scar tissue. But Rosenthal suspects that the newt's healing powers also lie hidden in mammalian genomes, and might be reawakened by appropriate tweaking. So for her, the mouse genome offers the promise of contributing to the emerging revolution in regenerative medicine. "My research remains hypothesis-driven," says Rosenthal, "but the sequencing of the mouse genome has opened a new world."

Rosenthal has already generated some intriguing results. She has, for example, shown that transgenic mice that overexpress the growth factor IGF-1 in their muscles develop Arnold Schwarzenegger-style biceps, due in part to an influx of stem cells from bone marrow that develop into new muscle cells. Along with other research groups, she has also found that IGF-1 can slow or reverse muscle loss in mouse models of degenerative diseases such as muscular dystrophy.

Rosenthal is now probing exactly how IGF-1 promotes regeneration. How does it



Nadia Rosenthal (above) says the mouse genome has “opened a new world”. Monica Justice has already used it to pinpoint recessive mutations.



cause stem cells to settle in muscle, and persuade them to differentiate into muscle cells? What genes are switched on and off as the stem cells evolve through different levels of commitment into muscle? To investigate, Rosenthal is using DNA microarrays to study the expression of hundreds of genes at a time. “But biology is more complicated than just finding out what genes are active at a particular stage,” she says. “We want to know all about the genes that regulate the gene, its regulators and their switches — and for this, only full genome information will help.”

In future, comparing the mouse genome with those of other species will be revealing, Rosenthal believes. “We can learn a lot by looking at the same genes in genomes of other species that are either bad or good regenerators,” she says. She also hopes to work with a remarkable mouse strain called MRL, bred at the Wistar Institute in Philadelphia, which can repair damage to its heart without scarring.

Rosenthal joined EMBL’s Italian outpost from Harvard University last year for the opportunity to work with the five other groups applying genetic and genomic approaches to the study of mouse biology. Her colleagues are similarly enthused about the opportunities that the mouse genome offers.

“We’ve really been waiting for the genome sequence,” says neurobiologist Liliana Minichiello, whose team is trying to understand the control mechanisms of the complex cellular signalling pathways triggered by nerve growth factors. “It is so easy now to identify the genes you are interested in from just a scrap of information.” ■

Allison Abbott

◆ www.embl-monterotondo.it

All mutants great and small

Hidden beneath a rolling lawn at Baylor College of Medicine in Houston, Texas, lies a subterranean colony of tens of thousands of mutant mice. The mice have been altered by a chemical that mutates their DNA, and their cage labels reflect their respective genetic hiccups. ‘Audrey Hepburn’ is a charming specimen with a perky, upturned nose. ‘Kojak’ has a striking pattern of hair loss. And the unfortunate ‘Wee Willie’ is under-endowed and infertile.

Monica Justice, director of Baylor’s Mouse Mutagenesis and Phenotyping Center for Developmental Defects, hopes that her growing menagerie of defective rodents will serve as models of human disease. But without a sequenced mouse genome, her project would be well-nigh impossible.

Take Wee Willie. To create him, a lab technician injected a male mouse with a chemical called *N*-ethyl-*N*-nitrosourea, which causes random, tiny mutations in sperm. The male’s offspring were then bred for several generations. Justice noticed that male mice in Wee Willie’s lineage sometimes had smaller penises than normal and couldn’t make sperm.

Nobody had ever seen a mutant like this, so Justice and her colleague Richard Behringer at the University of Texas MD Anderson Cancer Center in Houston decided to study it with Behringer’s postdoctoral fellow Andrew Pask. In the past, pinning down the mutation responsible for Wee Willie’s condition might have taken the best part of Pask’s career. But thanks to the mouse genome sequence, he and Baylor geneticist David Stockton were soon able to pinpoint the defect to a small region of chromosome 5.

Pask has found that Wee Willie’s defects mirror a human disease called idiopathic hypogonadotropic hypogonadism, in which patients fail to develop normal adult sexual characteristics. Wee Willie may become a useful model for studying this distressing disease, and it turns out that he is a type of mutant called a ‘hypomorph’ — his defective gene still functions, but doesn’t do its job well enough.

Established methods for making mutants, such as gene-knockout technology, can’t easily be used to make hypomorphs, a fact that vindicates Justice’s approach. There were plenty of doubters back in 1998, when she and her former colleague Allan Bradley set up the project. At that point, German and British teams had set up projects to screen for ‘dominant’ mutants, which disrupt development even if only one of a chromosome pair is affected. Justice’s screen is designed to find recessive mutants, which are harder to spot. Already, her project has identified some 200 mutants, and several other labs worldwide are now setting up recessive screens — including the Wellcome Trust Sanger Institute in Hinxton, near Cambridge, UK, now headed by Bradley (see page 512).

For Justice and her colleagues, the publication of the mouse genome is a time to reflect on a job well done — although much work still lies ahead. They are also looking back on a close brush with disaster. In June last year, thousands of mice at Baylor were lost in a catastrophic flood. Luckily, the designers of the new facility occupied by Justice’s mutants had the foresight to include watertight doors. ■

Erika Check

◆ www.mousegenome.bcm.tmc.edu/ENU/MutagenesisProj.asp

A forage in the junkyard

One of the main differences between the mouse and human genomes lies in the activity of 'junk' DNA sequences called retrotransposons. Carina Dennis considers what these sequences might be doing.

Wherever human garbage accumulates, mice will forage. But in genomics, the tables are being turned. For some geneticists, the most intriguing aspect of the mouse genome lies in its 'junk' — DNA sequences that don't code for proteins. The mouse, they believe, will prove a fertile hunting ground for researchers who want to understand why and how mammalian genomes accumulate this garbage, and whether it has any function.

Much of the junk is the work of transposons, genetic 'parasites' that have accumulated in mammalian genomes over millions of years by being copied into new genomic locations. Most are retrotransposons, which reproduce through an RNA intermediate, and so must use a reverse transcriptase enzyme to restore their original DNA sequence as they jump back into the genome.

The mouse genome is not especially full of such garbage — although at least 37.5% of it seems to be derived from retrotransposons¹, the proportion is similar in the human genome^{2,3}. But the activity of these retrotransposons differs greatly between the two genomes. In our evolutionary lineage, retrotransposon activity seems to have quietened down about 40 million years ago², and less than 100 are actively jumping around in any individual's genome today⁴. But in the mouse, about 3,000 of these parasites are thought to be still on the move⁵, on occasion disrupting host genes and affecting the creature's genome in ways that geneticists are keen to investigate. "I believe that some of these elements have been the most dynamic forces to sculpt the genome," says Sandra Martin, who works on retrotransposons at the University of Colorado Health Sciences Center in Denver.

Among the most common types of retrotransposon are LINEs — long interspersed elements — and their shorter counterparts, known as SINEs. LINEs can be up to 6,000 DNA base pairs long, and can enlist the host's gene transcription machinery to produce their own reverse transcriptase. SINEs are typically a mere 300 or so base pairs long, and lack their own reverse transcriptase, so are thought to co-opt the enzyme produced by LINEs to propagate themselves.



Say cheese: just one piece of 'junk' DNA can produce several coat colours in genetically identical mice.

The third major class, elements with long terminal repeats (LTRs), encode their own reverse transcriptase and bear similarities to retroviruses — indeed, they might be evolutionary relatives.

Retro chic

LINEs and SINEs cover almost 20% and just over 8% of the mouse genome, respectively; LTRs account for about 10% (ref. 1). All have been accumulating over tens of millions of years. "They are ancient fossils," says Haig Kazazian, a geneticist at the University of Pennsylvania School of Medicine in Philadelphia.

Whenever a retrotransposon lands in a host gene, it may alter the gene's function or disable it completely. And because many retrotransposons in the mouse are still active, they are thought to be responsible for about 10% of naturally occurring mutations that cause a noticeable change in characteristics⁴.

Retrotransposons don't just exert their effects by jumping into individual genes. In August, teams led by John Moran of the University of Michigan Medical School in Ann Arbor and Jef Boeke of the Johns Hopkins University School of Medicine in Baltimore, Maryland, demonstrated, in

cultured human cells, that retrotransposons can also cause the deletion of large blocks of DNA when they insert into a chromosome^{6,7}. Such deletions may have been important agents of evolutionary change.

Mutations and deletions caused by retrotransposons will, in many cases, be harmful. So most experts believe that the host genome is engaged in a constant war, fought through natural selection, to keep the genomic junkyard created by retrotransposons from spreading. The result is that the rubbish tends to accumulate in places where it doesn't get in the way and cause too much damage. Just as we might store our junk in a basement or closet, some corners of mammalian genomes are littered with retrotransposon remnants.

But other regions are spotless. Both the mouse and human genomes, for instance, seem to be meticulous about keeping the cluster of *Hox* genes that specify the body plan during embryonic development free from retrotransposons. Other similarly 'clean' regions might point researchers to genomic regions of crucial functional importance, suggests Dixie Mager, a geneticist at the British Columbia Cancer Research Centre's Terry Fox Laboratory in Vancouver, Canada. "Searching for regions where



Garbage men: Haig Kazazian (above) and Arian Smit are rooting through the rubbish dumps of the mouse genome for evolutionary clues.

retroelements are notably absent could offer insights," she suggests.

Strikingly, preliminary comparisons of the mouse and human genomes reveal that junk accumulates in similar places in both. Arian Smit of the Institute for Systems Biology in Seattle and his colleagues have compared blocks of sequence up to 500 kilobases long from mouse and human genomes, and found a similar pattern of density of retrotransposons, particularly SINEs, in corresponding regions¹. Because many elements jumped into the mouse and human genomes long after the two species diverged, they must have independently come to rest in similar positions.

Meanwhile, emerging evidence suggests that the host genome sometimes puts its resident retrotransposons to good use — much as we might dust off old junk and recycle it for a new purpose. Mart Speek of Tallinn Technical University in Estonia, for instance, has shown that LINEs can alter the expression of neighbouring host genes⁸. Again, most such changes are likely to be harmful, but in some cases, they may regulate gene activity in ways tolerated by the host.

"Young elements are interesting because they are still active and moving around the genome and possibly disrupting genes," says Mager. "But older elements are also interest-

ing because they have had time to become absorbed by the genome and possibly assume a regulatory function."

LINEs, for instance, seem to accumulate on the sex chromosomes of both human and mouse¹⁻³, and are particularly common in the regions of the X chromosome involved in shutting down gene expression in one of the two copies of the chromosome present in female mammals^{9,10}. Perhaps, some researchers speculate, LINEs have been recruited as part of the machinery responsible for this inactivation. But other experts think this unlikely, arguing that any regulatory role for retrotransposons will have arisen purely by chance. "There is no evidence of active recruitment," asserts Anthony Furano of the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland.

Line dancing

Irrespective of whether the mammalian genome has made use of some retrotransposons, researchers may be able to manipulate them to their advantage in the future. Kazazian and his colleagues have created a mouse model to study the movement of a human LINE, in which an active retrotransposon can be passed onto the next generation, inserting into a new genomic position as it does so¹¹. This will allow researchers to study retrotransposition and its consequences in a controlled way in living mammals for the first time.

Combine the potential of this model with the mouse's extensive evolutionary legacy of past retrotransposon activity, add the fact

that many of its retrotransposons are still moving around, and mice should become the ideal organisms in which to study the influence of retrotransposons on mammalian genomes. But it may be a couple more years before the full harvest can be reaped.

The current draft of the mouse genome was produced by a technique called whole-genome shotgun sequencing (see page 460). This method is notoriously poor at rendering the correct sequence of large expanses of junk DNA that contain the repetitive products of retrotransposon activity. Of the 3,000 potentially active retrotransposons thought to be in the mouse genome, only 400 — of which just 12 have their entire coding region intact — could be found in the assembled sequence¹. "It is disappointing, as important information has been lost," says Kazazian.

But the public consortium responsible for the draft mouse genome plans to finish the job by 2005, using a map-based method that should fill in this missing information. For retrotransposon aficionados, that holds the exciting prospect of being able to compare the influence of these elements on human and mouse genomes — and perhaps other species. A comparative sequencing effort along these lines is being spearheaded by Eric Green of the National Human Genome Research Institute in Bethesda, Maryland, who is generating large quantities of high-quality sequences from more than two dozen vertebrate species.

Other researchers want to compare retrotransposons between strains of mice, and even different mice within the same strain. Researchers led by Emma Whitelaw at the University of Sydney in Australia have already shown that a single retrotransposon can cause heritable differences in coat colour between mice that are genetically identical¹². Retrotransposons, it seems, may help to explain not only differences between species, but also a spectrum of traits between individuals — not a bad haul from a forage in the genome's junkyard. ■

Carina Dennis is Nature's Australasian correspondent.

1. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
2. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
3. Venter, J. C. et al. *Science* **291**, 1304–1351 (2001).
4. Ostertag, E. M. & Kazazian, H. H. Jr *Annu. Rev. Genet.* **35**, 501–538 (2001).
5. Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H. Jr *Genome Res.* **11**, 1677–1685 (2001).
6. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. *Cell* **110**, 315–325 (2002).
7. Symer, D. E. et al. *Cell* **110**, 327–338 (2002).
8. Speek, M. *Mol. Cell. Biol.* **21**, 1973–1985 (2001).
9. Bailer, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. *Proc. Natl. Acad. Sci. USA* **97**, 6634–6639 (2000).
10. Lyon, M. F. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
11. Ostertag, E. M. et al. *Nature Genet.* **32**, 655–660 (2002).
12. Morgan, H. D., Sutherland, H. G. E., Martin, D. I. K. & Whitelaw, E. *Nature Genet.* **23**, 314–318 (1999).

Piecing it all together

The whole-genome shotgun method has assembled a high-quality draft mouse sequence. Future projects will wed the shotgun's speed and economy to established, map-based methods, says Declan Butler.

"Tossed genome salad." With these and other disparaging words¹, leading members of the public human genome project last year derided the draft assembly of the human genome produced by Celera Genomics of Rockville, Maryland². Without access to sequence³ and mapping⁴ information produced by the public project, its leaders claimed, Celera would have been nowhere⁵.

But these arguments, with their culinary images about who had the superior recipe for sequencing, obscured the fact that future mammalian genome projects would incorporate ingredients of both team's approaches. Despite the continuing debate over whether it succeeded with the human genome, Celera's approach, dubbed the whole-genome shotgun (WGS), was adopted by the publicly funded researchers for their encore, the draft mouse genome. The resulting sequence⁶ surpasses the quality of last year's public draft human genome³.

"The lesson of the mouse assembly is that it is possible to get a good draft using a whole-genome shotgun," says George Weinstock, co-director of the Baylor College of Medicine Human Genome Sequencing Center in Houston, Texas.

The gold standard for sequencing a mammalian genome is the clone-by-clone method, used to produce the public version of the human genome. Here, the genome is chopped into chunks up to several hundred thousand base pairs long, and cloned into bacterial artificial chromosomes (BACs). Each BAC is sequenced by shattering its cloned DNA into smaller fragments and then reassembling the sequence using computer algorithms that match the overlapping ends. Because BACs can be mapped onto the genome's chromosomes by looking for markers called sequence-tagged sites (STSs), the entire sequence is gradually built up. The downside is the cost, time and effort of constructing and sequencing in depth the library of 20,000 or more BACs that is needed to map the entire genome.

The WGS approach avoids this hassle. Rather than producing a BAC map, the genome is instead blasted into millions of fragments, which are sequenced and reassem-

bled to produce a series of sequence 'scaffolds'. These can then be mapped directly onto the chromosomes using STSs. The problem is that mammalian genomes contain millions of repeated sequences. For the assembly algorithms, the challenge is similar to completing a jigsaw that comprises mostly blue sky.

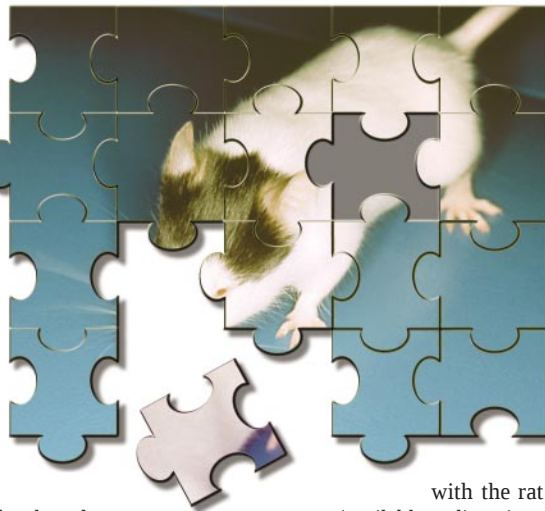
The mouse assembly's success stems from several factors, including the availability of sophisticated public-domain assembly software. At the Whitehead Institute Center for Genome Research at the Massachusetts Institute of Technology, computational biologists have developed a package called ARACHNE, and the Wellcome Trust Sanger Institute at Hinxton, near Cambridge, UK, has produced an equivalent known as Phusion. "They use a whole bunch of computational tricks," says Eric Lander, director of the Whitehead genome centre.

Squeaky clean

Improvements in sequencing technology, giving longer and more accurate sequence reads, have also allowed the mouse genome team to make abundant use of 'mate pairs' — sequences from either end of a DNA segment of known length. If mate pairs subsequently end up the wrong distance apart within the final assembly, it suggests a problem.

Most significantly, perhaps, it has been possible to nail the WGS scaffolds to excellent genetic⁷ and physical⁸ maps of the mouse genome. The latter was created in part by using the human genome as a 'cheat sheet'. Mice genes align well with human ones, and large blocks exist in which gene order is the same in both genomes. The BACs will now be sequenced in detail over the next couple of years to produce a clean, finished sequence in which gaps are closed and errors corrected.

Although the mouse assembly confirms the utility of the WGS, in itself it says little about the optimal hybrid strategy. But along



with the rat genome (available online since 25 November at www.hgsc.bcm.tmc.edu), computational biologists now have enough WGS and clone-by-clone data to try out assemblies using different combinations of each. The best strategy for future projects will depend on the organism, whether a draft or finished sequence is needed, and species' genetic variability.

Looking further ahead, similarities in gene order across mammalian genomes should enable researchers to sequence BAC clones lightly, before comparing them to the human and mouse sequences to construct rough physical maps. With these in hand, it should be possible for researchers interested in comparative genomics to dip into sequence-specific regions of interest in several different species.

The WGS, meanwhile, seems set to become the method of choice for producing rough drafts of large genomes. Clone-by-clone sequencing will remain the *haute cuisine* for finishing genomes, but the WGS's fast food is now firmly on the menu. ■

Declan Butler is Nature's European correspondent.

1. Butler, D. *Nature* **409**, 747–748 (2001).
2. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
3. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
4. International Human Genome Mapping Consortium *Nature* **409**, 934–941 (2001).
5. Waterston, R. H., Lander, E. S. & Sulston, J. E. *Proc. Natl Acad. Sci. USA* **99**, 3712–3716 (2002).

Towards taxonomy's 'glorious revolution'

Taxonomy is a triumph of modern science — but its products could still be improved.

Sir — My commentary “Challenges for taxonomy” (*Nature* **417**, 17–19; 2002), in which I argued for unitary, web-based taxonomies, has stimulated a lively debate. In Correspondence, K. Thiele and D. Yeates (*Nature* **419**, 337; 2002) make the important point that, unlike stars or genes, species and other taxa are hypotheses, not facts. A taxonomic name is thus a shorthand for this hypothesis, but it is also the key that indexes all the biological information on that taxon.

I agree wholeheartedly that any unitary taxonomy must recognize this duality. That is why I suggested that while the ‘current web revision’ would present only a single taxonomy for the user community, any alternative hypothesis would by right be lodged on the website and possibly enter a future revision. I believe a unitary taxonomy can be authoritative without being authoritarian.

In another Correspondence, I was disappointed that the high command of London's Natural History Museum found so little merit in these ideas (S. Knapp *et al.*, *Nature* **419**, 559; 2002), but I fear I must have explained them poorly. I do not seek to “throw out the past mechanisms of doing systematics and begin anew in a revolutionary ‘brave new world’”; nor to do away with the order and stability conferred by type specimens, nor to belittle the major taxonomic web projects already in existence. I think we should tinker as little as possible with the process of taxonomy, although my proposals for the products of taxonomy are more radical. However, exploring whether they might work for one group or for a few groups is only a small, reversible step.

Taxonomy as currently practised works and we need much more of it. It is one of

the triumphs of modern science. But unless it looks to its end-users and provides them with products that they will both use and actively campaign for (and I believe this can be done while maintaining the subject's core methodology), I worry deeply that it will lose out in the intense competition in the sciences for funds.

I wrote that one of the advantages of unitary taxonomies is that they are “evolutionary, not revolutionary”, and so was amused to be labelled a revolutionary. But there are revolutions and revolutions, and while taxonomy does not need the 1789 French version with its ensuing reign of terror, perhaps it does need something like England's ‘Glorious Revolution’ of 1688.

H. C. J. Godfray

NERC Centre for Population Biology,
Imperial College at Silwood Park,
Ascot, Berkshire SL5 7PY, UK

US embryo rules support ‘right-to-life’ agenda

Sir — I believe that the crisis facing embryo research in the United States is a great deal more serious than your news item “US biologists wary of move to view embryos as human beings” (*Nature* **420**, 3–4; 2002) portrayed it. The Bush administration's clear intent is to establish a legal precedent for declaring embryos to be people from the moment of conception. If this effort succeeds, all research on embryos in the United States could cease.

Assurances that the changes are intended only to protect pregnant women are obvious falsehoods. If that were the case, the new remit would emphasize “pregnant women”, not “pregnant women, embryos, and fetuses”. These assurances have been made by the deputy assistant secretary of the Health and Human Services (HHS) department, Arthur Lawrence, reported in *The Washington Post* (30 October 2002, page A1), as well as by HHS spokesman Bill Hall, as reported in your News story. Any doubts about the political motivation of these changes should be settled by the Bush administration's stated intention, also reported in *The Washington Post's* article, to place on the new HHS committee Mildred Jefferson — a founder and past president of the National Right to Life Committee.

Adherents of the US ‘right to life’ movement have made backdoor attempts at legislation before. A proposal to declare conception as the beginning of legal

personhood by constitutional amendment has been floating around for many years. Even the amendment's supporters concede that it has no chance of ratification, however, so attempts proliferate to obtain by executive fiat that which is unobtainable by straightforward appeal to US citizens. Given the results of the 5 November elections, we can expect more such attempts in the near future.

The biomedical community and patients' advocacy groups should harbour no illusions about what is at stake. The goal of certain elements in US politics is to have balls of undifferentiated cells legally defined as people. From that beginning, the argument that a ball of cells cannot give informed consent, and therefore cannot be a test subject, is a short and inescapable step away.

Gordon G. Cash

Address supplied

Would you give up your grant for sustainability?

Sir — Your Opinion editorial “Leadership at Johannesburg” (*Nature* **418**, 803; 2002) reveals delusions about the sustainable-development summit. Most scientists care little about sustainability, and very few work towards achieving it. I have tried to excite the geoscience community's interest, for example (*Geoscientist* **11**, 1 (15) and 5 (11); 2001), but elicited only a meagre response.

You write of poverty alleviation, improved (renewable) energy generation

and better water supplies in developing countries. Less poverty leads to greater consumption of goods and services, of energy, water, land and so on. That is not sustainable development. The Correspondence by E. D. G. Fraser and W. Mabee (*Nature* **418**, 817; 2002) correctly predicted that the summit would fail, but shed no new light on the issues.

Which of the following do readers consider compatible with sustainable development? Expanding airports and road networks, growing car and aviation industries (including motor racing)? The international arms trade? Logging of rainforests? Trapping and shooting of endangered birds and animals? Overfishing and progressive pollution of the oceans? Inability to police environmental legislation anywhere in the world? Corruption at all levels of government throughout the world?

In a piece entitled “Sustainable development unsustainable” (*Nature* **374**, 305; 1995), John Maddox wrote an approving review of Wilfred Beckerman's book *Small is Stupid*. Beckerman was wrong to imply that Earth's resources are not finite, but he and Maddox were right to point out that we need economic growth.

Can anyone name a country, rich or poor, where sustainable development is a central plank of government policy? Sustainable development means living in equilibrium with natural cycles. That means reduced consumption, which very few people want. Any attempt to wean industrial civilization off fossil fuels would lead to instant global recession and

economic decline. Without economic growth there would be no profits, hence no taxes, no public services, no pensions or state benefits — and no research grants!

John Wright

*Department of Earth Sciences, Open University,
Milton Keynes MK7 6AA, UK*

DNA committee is model for bioterrorism debate

Sir — There is much debate about how to protect the public from bioterrorism while maintaining the open exchange of biomedical research findings (see, for example, *Nature* **419**, 99 and 769; 2002). There have been several calls by scientists and science journalists in the media for a 'summit conference' to discuss these concerns and to establish a US national forum to guide federal policy on scientific research with possible bioterrorism applications.

Fortunately, a highly effective model already exists. The 1975 Asilomar conference led to the establishment of the NIH Recombinant DNA Advisory Committee (RAC) to provide guidance and oversight of research using recombinant DNA (see, for example, R. M. Atlas *Science* **298**, 753–754; 2002).

Piecemeal regulation will harm the advancement of science and wreak havoc on national security. Already the US Office of Management and Budget is planning to draft rules on the publication of some federal research classified as "sensitive homeland security information", but these rules will apply only to US government laboratories. In testimony before the House of Representatives Science Committee, several university officials said the proposed new rules would not serve the best interests of science.

A broadly based committee modelled on the RAC should be established now under the aegis of the NIH to develop guidelines for biomedical research with possible bioterrorism applications. A comparable committee should be formed in the US Department of Agriculture to deal with agricultural research. An interagency coordinating committee would be needed for supervision on security concerns related to government-sponsored research.

An agenda for an RAC-like committee has been laid out by questions raised since last year's anthrax attacks. First, should access to genome databases be restricted? Would such restrictions prevent bioterrorism? More than 60 microbial genomes have already been deposited in publicly available genome databases and work is actively progressing on more than

150 more, as well as on several parasitic, fungal, plant and insect genomes.

Second, should scientists and scientific publishers practise self-censorship? For example, we now know that live poliovirus can be synthesized and that mail-order companies will provide the building-blocks necessary to reconstruct the virus in a laboratory. What, if any, restrictions on the publication of scientific findings are legitimate?

Third, should the training of scientists in the United States, whether US citizens or foreign nationals, be restricted? New US legislation requires universities and laboratories dealing with "select agents" to be registered with the US Department of Health and Human Services or the US Department of Agriculture and requires individual scientists to submit to background checks by the Department of Justice. Will an increasingly perilous legal landscape discourage legitimate scientists from engaging in research on infectious diseases?

The RAC model has served us well in the past; we should use it as the starting point for addressing these key policy questions to achieve effective and optimal guidance for the future.

Joseph G. Perpich

*JG Perpich, LLC, 7315 Wisconsin Avenue, Suite 500
East, Bethesda, Maryland 20814, USA*

Modelling a new angle on understanding cancer

Sir — There was a striking absence of discussion of quantitative methods in the recent interesting theoretical debate on the timing and nature of mutations controlling acquisition of the metastatic phenotype in invasive cancers (R. Bernards and R. A. Weinberg *Nature* **418**, 823; 2002 and Correspondence *Nature* **419**, 559–560; 2002). Without the discipline imposed by mathematical rigour, the arguments have to rely on intuitive reasoning and small pieces of data singled out from the vast extant literature. The reader is unable to independently evaluate the hypotheses or understand them in the context of all available experimental information or of theoretical models that organize information in integrally related phenomena such as carcinogenesis and tumour invasion. Although this state of affairs is normal in some disciplines such as tumour research, it would be unthinkable in the physical sciences.

Medical investigators often eschew mathematical models as too "simplistic" for the dynamics governing processes such as those in tumour biology. Instead, the general underlying principles in complex

biological processes are expected to become apparent, as if by magic, once sufficient data are accumulated. Not surprisingly, this passive approach has failed to yield a comprehensive theoretical framework of understanding carcinogenesis, tumour invasion and metastatic behaviour. Clinical therapy thus remains largely empirical, based more on trial and error than a comprehensive understanding of biological first principles.

In most other scientific disciplines, there is an active search for broad generalizations through hypotheses framed in mathematical models that can be examined for internal consistency and compatibility with extant data. Predictions from these models can be tested by experiment and revised when necessary. Surely the success of this integrative approach, combining quantitative theoretical methods and experiment in highly complex systems in physics (including biophysics), chemistry, engineering and biological disciplines, such as ecology or structural biology, should motivate its enthusiastic application to other nonlinear biological processes such as carcinogenesis and tumour invasion.

In the absence of consistent application of rigorous mathematical models, theoretical medicine will largely remain empirical, phenomenological and anecdotal, successful only in linear systems that can be defined by a single experiment or a few experiments. Until the quantitative methods of the physical sciences are applied, many difficult and clinically important problems in tumour biology, including the dynamics that give rise to the metastatic phenotype, will remain unresolved.

Robert A. Gatenby*, Philip Maini†

** Departments of Radiology and Applied Mathematics, University of Arizona, 1501 Campbell Avenue, Tucson, Arizona 85724, USA
† Centre for Mathematical Biology, Mathematical Institute, 24–29 St Giles', Oxford University, Oxford OX1 3LB, UK*

Deserted by our geographical sense

Sir — Your News in Brief on the Atacama Large Millimeter Array in Chile (*Nature* **419**, 870; 2002) sets a new record for imprecision. Chile is a narrow ribbon, with an average width of just 180 km and an eastern frontier of about 6,000 km of virtually constant longitude. Hence, referring to the observatory as located: "in the Atacama desert in the east of the country" is definitely too indefinite for anyone wanting to find it.

Gustavo Artega

Département de Chimie et Biochimie, Laurentian University, Sudbury, Ontario P3E 2C6, Canada

The quiet man of physics

Who is the only physicist to have won two Nobel prizes?

True Genius: The Life and Science of John Bardeen

by Lillian Hoddeson & Vicki Daitch

Joseph Henry Press: 2002. 482 pp. \$27.95

P. W. Anderson

John Bardeen was an extremely quiet man. An anecdote in this book avers that when he was selling his house in Summit, New Jersey, the prospective buyer was so disconcerted by Bardeen's silence that he raised his bid by many thousands of dollars while waiting for him to speak. As an old friend of Bardeen's, I don't find that at all implausible.

Nonetheless, this quiet man led the way in two earth-shattering developments: with Walter Brattain he devised the first working semiconductor amplifier, jump-starting the information revolution; and with two young associates, he solved the 46-year-old puzzle of superconductivity, with repercussions not just in that field but in fundamental aspects of nuclear and elementary high-energy physics. He also helped to plan the research laboratories of the giant Xerox Corporation, and was a friend and consultant to the founder of Sony. By the way, he is the only person ever to have won two Nobel prizes in physics.

Still, as Lillian Hoddeson and Vicki Daitch bewail in this book, to most of the world he is "John who?". They intend to start remedying this situation. The second theme of the book is also the source of the title: they want to emphasize that it is possible to be a "true genius" without being bohemian, neurotic or particularly difficult to get along with — and that perhaps most geniuses do not fit the popular stereotype.

Of course, the biographer's task is much harder if the subject is not eccentric in any spectacular way. Even Bardeen's precocity — he entered college at 15 — did not seem to impair his ability to get on well with his peers. A few years of indecision as to his natural calling ended happily in 1933, at the age of 25, with admission to Princeton University's graduate school, where he joined the fortunate little band of extraordinary students of Eugene Wigner who founded the modern quantitative theory of solids.

From Princeton he went on to be admitted into the elite Society of Fellows at Harvard, which gave him time to attempt several of the toughest problems in his field. He finally achieved a professorship at Minnesota University, and married Jane Maxwell after three years of courtship — he had the old-fashioned determination to have a positive cashflow before marrying. Jane was attractive, universally beloved, patient and, happily, chatty. Soon called into war work, Bardeen



Radio days: (left to right) John Bardeen, William Shockley and Walter Brattain invented the transistor.

rose to head a group of 90 engineers and scientists at the uncomfortable and crowded Naval Ordinance Lab in Washington.

The drama in his life was largely compressed into the next dozen years. A generous offer drew him to Bell Labs, where a group was being assembled under Bill Shockley to attempt to create an amplifier based on wartime advances in semiconductor science. The story of that achievement (by December 1947) is told more fully in Hoddeson's previous book *Crystal Fire*, written with Michael Riordan (Norton, 1997). But in *True Genius*, Hoddeson and Daitch recount the full story of the personal conflict with Shockley that ensued when the latter used his position to isolate and overwhelm the two original inventors. Perhaps, it must be said, this had lasting benefits for later developments in semiconductors, because it drove Bardeen to work on advancing the science generally, while Shockley pursued a line that eventually petered out.

By 1951, Bardeen had been lured away to the University of Illinois by old friend and fellow Wigner student Fred Seitz. There he not only fostered an independent semiconductor-engineering group with Nick Holonyak, but built a theoretical-physics group within the already thriving condensed-matter physics enterprise.

He was at last able to indulge his obsession with the puzzle of superconductivity, left over from his Harvard days and reactivated in his final months at Bell. Using again his obstinacy and the careful assembling and critical interpretation of experimental data

that had carried him so quickly to the transistor, Bardeen and David Pines soon formalized the crucial interaction. By the end of 1956 (with a slight interruption to pick up the Nobel Prize for the transistor), Bardeen, his student Bob Schrieffer and postdoc Leon Cooper had solved that problem as well. The resulting paper (*Physics Review* **108**, 1175; 1957) deserves a place among the all-time classics of science. The intellectual ramifications of their central hypothesis — known as broken gauge symmetry — pervade all of physics, and have underpinned at least four other Nobel prizes to nine individuals. There will doubtless be more to come.

After that, the rest of Bardeen's life may seem like an anticlimax, even though it included key roles in the creation of the Sony Corporation and the Xerox Palo Alto Research Center laboratories, and high office and influential advisory roles in the scientific establishment. Two well-written and well-researched chapters are devoted to scientific controversies in which Bardeen became embroiled in later life, in several of which it is hard in retrospect to support his side of the matter. He may have been quiet but he was not self-effacing, and was quite capable of throwing his weight around in such controversies. But when the dust settled he could be quietly generous to his opponents, and was instrumental in the later Nobel Prize awards to Brian Josephson and probably to me.

To return to the genius thing. Is it possible to be a genius when you are sometimes wrong? (Personal glamour seems to have

BETTMANN/CORBIS

little to do with it one way or the other, in science as in the arts.) Actually, few scientific geniuses are even close to infallible, and some can be wrongheaded indeed — look at Linus Pauling and vitamin C, or Julian Schwinger and cold fusion. Bardeen cracked a problem that dozens of the greatest minds of physics had failed at — isn't that enough?

There are, almost inevitably, glitches in the details in this book, but the authors' admiration and affection for their subject illuminates the biography. At the same time, they bring readers with varied levels of expertise to a real understanding of the complex workings of science as they are actually experienced by those of us who do it. Will the book appeal to the mass market and thus make a dent in the "John who?" problem? I'm not convinced, but perhaps it should. ■

P. W. Anderson is in the Department of Physics, Princeton University, Princeton, New Jersey 08544-0708, USA.

Keeping your feet in a moving field

Earthshaking Science: What We Know (and Don't Know) about Earthquakes

by Susan Elizabeth Hough

Princeton University Press: 2002. 272 pp. \$24.95, £17.95

Gregory C. Beroza

Seismologists have been grouped, unfavourably, with economists in that they are both great at telling you why an event happens after it has happened. This sentiment reflects frustration with our limited knowledge of earthquake behaviour. Why don't we understand earthquakes better? And why can't we predict them?

Large earthquakes occur infrequently, and the processes that are the key to understanding them are removed from our sight by miles of opaque material. Moreover, the sudden rupture of a fault in an earthquake is a nonlinear process that occurs within materials that vary strongly in their properties and appear to be complex at all spatial scales. Given these challenges, perhaps it is not surprising that we lack a comprehensive understanding of earthquake behaviour.

A more optimistic view of our limited understanding is that the science of earthquakes is not mature. Almost every new well-recorded earthquake seems to have at least some surprising aspect. We are still on the steep part of the learning curve.

Earthshaking Science takes on the difficult task of reviewing the state of earthquake science at a time when the field is evolving rapidly. Its author, Susan Hough, has done an admirable job of clearly and accurately

illuminating the boundary between our knowledge and our ignorance. In the process, she outlines some of the major outstanding problems in the field and provides a balanced and insightful view of them from several sides. Time and again as I read this book I thought to myself that she had advocated a particular position too strongly, only to find on the following page that she would offer an equally strong counterargument.

Hough, who has experience of both earthquake research and public outreach, has written a book that is accessible to readers in other disciplines and to a non-technical audience, but provides enough thoughtful commentary and perspectives to hold the attention of specialists.

An important part of any general book on earthquakes is how it treats short-term earthquake prediction — a topic that is central to seismology and foremost in the minds of non-seismologists. In this book there is not a great deal of material on earthquake prediction, which might disappoint readers outside the field. This de-emphasis is not surprising, though, given the lack of success to date. Instead, the discussion focuses on whether or not earthquakes are predictable even in theory (see *Nature* web debate; <http://www.nature.com/nature/debates>). As elsewhere, Hough delivers an even-handed and up-to-date treatment of both sides of the issue.

The book excels in its treatment of the prediction of potentially damaging strong ground motion in the near field of an earthquake. The ability to predict strong ground motion is arguably more important than the ability to predict earthquakes. Seismologists use ground-motion prediction in the form of probabilistic seismic hazard analysis (PSHA) to characterize earthquake risk.

PSHA, as used in building codes, for example, is an estimate of the distribution of ground motion with a 10% probability of being exceeded over a 50-year time interval. A lot goes into this estimate, and this book explains it clearly. To my knowledge there is no other accessible treatment of this topic. The test that precariously balanced rocks offer of the validity of PSHA is a nice example of the currency of the material in this book.

Earthshaking Science is not a textbook or a coffee-table book. But it is a readable tour of many key aspects of earthquake science. The author focuses on California but also covers the poorly understood earthquakes in the central and eastern United States. What's missing is a thorough treatment of the many earthquakes in other tectonic environments. One can hardly fault the author for this, because it would require a considerably larger book. Moreover, much of the research knowledge needed to write such a book at the same level does not yet exist. But it soon will.

Recently deployed modern seismic and geodetic monitoring networks in Japan, as



Shock result: earthquakes do enormous damage but are extremely difficult to predict.

well as parallel efforts being contemplated in the United States, have led to the discovery of new phenomena. In the past year, discoveries of large aseismic transients and what appears to be an entirely new type of seismic event deep under Japan are changing our views of fundamental earthquake processes. It is my strong expectation that the author will have a lot of new material to incorporate into the second edition. ■

Gregory C. Beroza is in the Department of Geophysics, Stanford University, Stanford, California 94305-2215, USA.

More on earthquakes

The Mechanics of Earthquakes and Faulting, 2nd edn

by Christopher H. Scholz

Cambridge University Press, \$130, £90 (hbk); \$48, £32.95 (pbk)

Fertile ground for politics

Autarkie und Ostexpansion: Pflanzenzucht und Agrarforschung im Nationalsozialismus

edited by Susanne Heim

Wallstein Verlag: 2002. 312 pp. E20

Wissenschaften und Wissenschaftspolitik: Bestandsaufnahmen zu Formationen, Brüchen und Kontinuitäten im Deutschland des 20. Jahrhunderts

edited by Rüdiger vom Bruch & Brigitte Kaderas

Franz Steiner Verlag: 2002. 476 pp. E96

Marco Finetti

Germany's major scientific organizations are finally examining their past during national socialism. For years the involvement of the sciences, and in particular of the research organizations, in the Third Reich was pursued only by undaunted PhD students or

journalists. But now both the Deutsche Forschungsgemeinschaft (DFG), Germany's main research funding agency, and the Max-Planck-Gesellschaft (MPG) are officially preoccupying themselves with the subject. Under external pressure, as well as from an internal sense of judiciousness, they have commissioned independent historians to reappraise their activities, or those of their predecessor organizations, during the years from 1933 to 1945, and in the early years of the Federal Republic of Germany. And there is much more to be brought to light.

The superb anthology *Autarkie und Ostexpansion* looks at the way in which German plant geneticists in several institutes belonging to the Kaiser-Wilhelm-Gesellschaft, the predecessor of the Max-Planck-Gesellschaft, surrendered themselves to the National Socialist rulers. The 11 contributions in this edited volume describe a science that represented particularly fertile ground for National Socialist ideology and politics.

Agricultural self-sufficiency had been on the political agenda of plant scientists

long before Hitler came to power. Nazi agricultural policy, however, went one decisive step further: it linked autarchy with expansion. As the book appropriately puts it, this appeared to the plant geneticists to be "a welcome framework in which long-planned research programmes could finally be carried out".

The zealotry with which the scientists proceeded and the ultimate outcome are illustrated by the example of Wilhelm Rudorf. Several of the anthology's articles examine his double career before and after 1945. Somewhat mediocre as a scientist but loyal as a National Socialist, Rudorf was appointed director of the Kaiser Wilhelm Institute for Cultivation Research in Müncheberg near Berlin in 1935, at the explicit request of the Nazi rulers but in opposition to the experts.

The ambitious Rudorf showed his supporters his gratitude. In 1937, long before Hitler's invasion of Poland and the Soviet Union, he called on German plant cultivators "to develop or improve useful plants

which will make it possible to populate more densely the entire northeastern and eastern territories and other border areas". And that was not all — from 1943, at the personal request of Heinrich Himmler, Rudorf helped to set up a centre for the cultivation of natural rubber in the grounds of the Auschwitz extermination camp — an impressive example of cooperation between the Kaiser-Wilhelm-Gesellschaft and Himmler's élite corps, the SS.

This cooperation with the Third Reich did not stop Rudorf from pursuing his career after 1945. He remained director until 1961, and until his death in 1969 was a member of the newly founded Max Planck Institute for Cultivation Research, initially in Hameln, later in Köln-Vogelsang, where the internationally renowned research institute still has its headquarters today. As director, Rudorf even prevented his colleague Max Ufer from resuming his work at the institute. Ufer had been dismissed from the Kaiser-Wilhelm-Gesellschaft in 1933 because he was married to a Jew. When Ufer asked to be reappointed in 1952, Rudorf agreed on condition that Ufer should not reside on the grounds of the institute because of his Jewish wife. Not surprisingly, Ufer turned down this supremely contemptible offer.

In many ways, *Autarkie und Ostexpansion* is both enlightening and shaming, and with its publication the Max-Planck-Gesellschaft has taken a positive step forward in reappraising its past. The same, unfortunately, cannot be said of Germany's biggest research organization, the DFG. Three years ago, it presented the results of an initial investigation into its own activities during the time of the Weimar Republic and the Third Reich. But the study, prepared by Frankfurt historian Notker Hammerstein, played down the DFG's involvement in the Nazi regime and met with criticism even from within the ranks of the DFG (see *Nature* **402**, 461; 1999).

A second attempt has now been made in the anthology *Wissenschaften und Wissenschaftspolitik*, produced at the initiative of the DFG. But this, too, is disappointing. The collection of almost 50 contributions covers a vast range of topics, including changes in the philosophy of modern physics before the First World War, 'big science' and research under National Socialism, and sociology in the Third Reich and in West Germany, to name just three. Most of the articles only scratch the surface, particularly regarding the role of the DFG in these areas. There is no doubt that the DFG, and above all its president, Ernst-Ludwig Winnacker, is serious in its attempts to reappraise the organization's past. But one could have certainly wished for a more enlightening choice of authors and subjects. ■

Marco Finetti is a member of the editorial staff at the *Süddeutsche Zeitung* in Düsseldorf, Germany.



Mountains of fire

Few sights on Earth are more spectacular, or more deadly, than a volcanic eruption. This dramatic image shows an eruptive cone and its lava flow at the Piton de la Fournaise volcano on the island of Réunion in the Indian Ocean.

It is one of 170 photographs taken by the award-winning lensman Philippe Bourseiller that are spectacularly reproduced at a giant size in *Volcanoes* (Harry N. Abrams, \$49.95), which includes text by Jacques Durieux.

Ultraviolet upset

Henry C. Kapteyn and Todd Ditmire

The free-electron laser will be a source of intense, short-wavelength radiation for a range of applications, including biological imaging. The first results from a prototype have already thrown up a surprise.

On page 482 of this issue, Wabnitz *et al.*¹ report on the first experiments with the free-electron laser at the TESLA test facility at DESY, Germany's high-energy physics research centre in Hamburg. They have seen unexpectedly strong absorption of the laser radiation by atomic clusters, resulting in ejection of high-energy ions. Such a highly nonlinear optical interaction between light and matter at ultraviolet wavelengths has never been seen before, and these fascinating results show the potential of this new class of intense short-wavelength lasers for scientific investigation.

Understanding the interaction of light with matter has been a central theme of physics over the past century, starting with Planck's development of the concept of the photon and the inception of quantum theory. Since then, the continuing advance of laser technologies in the 1980s and 1990s has made it possible to explore regimes of intense interaction, leading to technologies such as particle accelerators based on compact lasers², advanced concepts for fusion energy³ and methods for creating ultrafast light pulses at wavelengths ranging from the far infrared to γ -rays⁴.

The use of low-intensity ultraviolet light and X-rays is a staple of modern chemical, biochemical and materials science, and the physics of how this radiation interacts with matter is well understood. Light in the hard-X-ray regime, corresponding to very high-energy photons, has a wavelength comparable to the spacing between atoms in a solid, making X-ray diffraction techniques possible. At wavelengths in the far-ultraviolet region ('vacuum ultraviolet', or VUV), the energy of a photon becomes large enough for single photons to ionize matter. But light sources capable of producing intense light at these short wavelengths have not been available, and consequently the interaction of such light with matter remains a relatively unexplored area.

Free-electron lasers (FELs) have been used in the past for mid- and far-infrared light generation. They are also a promising way to generate intense short-wavelength light,

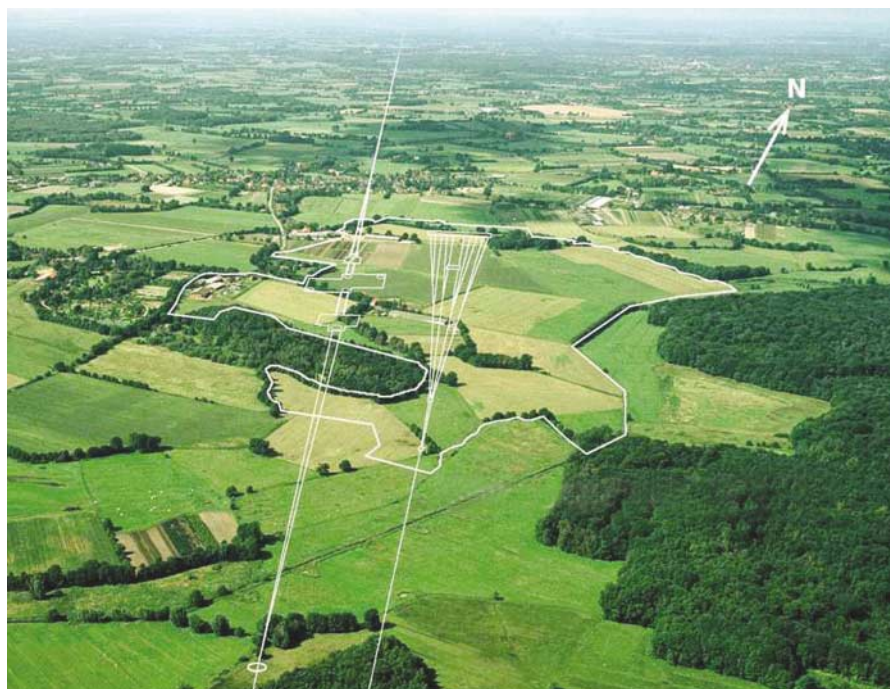


Figure 1 The planned layout of the TESLA free-electron laser facility. The TESLA project comprises an electron–positron linear collider, stretching 33 km north from the DESY laboratory in Hamburg, Germany (the white line to the left shows the proposed route of the accelerator). The accelerated electrons will also power a high-intensity free-electron laser, whose beam will be fanned out to several experimental areas (centre) at Ellerhoop, 16 km from DESY.

because the simplicity of the devices makes it possible to predict how to scale them to generate shorter wavelengths. In the FEL, an intense, energetic electron beam is made to pass through a periodic magnetic field, which causes the electrons to oscillate, and as they do so they emit light. The electron beam and the generated light interact cooperatively. This has the advantage, compared with the synchrotron-radiation light sources that have been used for decades, that in the FEL the light is produced in a coherent, tightly collimated beam that can be focused to high intensity.

The accelerator community has developed a number of proposals for short-wavelength FELs that could eventually be used to generate intense pulses of hard X-rays, just femtoseconds (10^{-15} seconds) in duration⁵. The potential impact of these X-ray lasers is enormous. For example, such a source could take 'snapshot' pictures of the atomic structure of a single biomolecule. Existing X-ray crystallography techniques work only for biomolecules that can be crystallized, but an intense FEL X-ray pulse

would make it possible to pass enough photons through a single molecule to effectively take a photo before the molecule explodes from the energy it has absorbed⁶.

Two large-scale FEL projects are currently in the planning stages — the Linac Coherent Light Source project at Stanford and the TESLA project in Hamburg. TESLA, with a proposed multi-billion-dollar particle accelerator and FEL facility (Fig. 1), has progressed furthest. The TESLA test facility is a smaller, proof-of-principle project that has begun operation at far-ultraviolet wavelengths, and it is results from this device that Wabnitz *et al.*¹ now present.

The authors illuminated xenon clusters with high-intensity FEL pulses of wavelength 98 nm — around a fifth of the wavelength of visible light. In this regime, a single photon has just enough energy to ionize a xenon (Xe) atom and remove a single electron. But Wabnitz *et al.* found that when nanometre-sized clusters of Xe atoms were irradiated with intense light from the FEL, highly charged energetic ions were produced. It

Further News and Views coverage appears on pages 515–519, with three articles discussing publication of the mouse genome sequence and related research.

seems that atoms in the cluster absorb an average of several dozen photons each, rather than the single photon that an isolated atom can absorb. Charge states up to Xe^{8+} were produced.

Enhanced light absorption and ion production in clusters is not new — experiments using high-intensity infrared lasers have observed both of these phenomena. This cluster absorption process can even lead to temperatures hot enough to induce nuclear fusion⁷. But the well-understood physics that describes the interaction of a cluster with a longer-wavelength, near-infrared laser pulse does not seem to work for the DESY experiments with VUV pulses.

Usually, high-energy ions are generated from an atom cluster in a 'Coulomb explosion', which occurs when the positively charged ions repel each other after the laser pulse has removed some or all of the electrons from the cluster. It is surprising that Wabnitz *et al.* see a Coulomb explosion in the case of VUV illumination, and moreover that it occurs at much lower illumination intensities than for infrared light. The significance of this effect can be estimated by calculating the 'ponderomotive energy' of free electrons — the kinetic energy of their oscillatory motion caused by the electric field of the laser light. For the case of infrared illumination with laser power $10^{16} \text{ W cm}^{-2}$, the ponderomotive energy is more than 500 electron volts. But in the DESY experiment the ponderomotive energy is less than 0.1 electron volts: the electrons' inertia makes them more sluggish in reacting to the higher-frequency light.

So electron motion driven by the laser's electric field should not be a factor in the new experiment, and direct laser-ejection of the electrons was not expected. But the clear evidence of a Coulomb explosion for VUV wavelengths raises questions about the physics of the absorption interaction. Energy is efficiently absorbed, but it is unclear just how this happens. As is well known in infrared irradiation of clusters, once atoms in the cluster become ionized, electrons can absorb energy in the same way that any plasma can absorb energy — by frequent random collisions between electrons. This process will begin at much lower intensities in the VUV case, as free electrons are immediately generated from the clusters by direct absorption of light. Unfortunately, straightforward models of the collisional absorption process underestimate the energy absorbed in the DESY experiment by a factor of ten.

The findings of Wabnitz *et al.*¹ do not suggest that the laws of physics are incorrect, but they do indicate that our theoretical estimates of collision-induced absorption, which work well in extended media, are not strictly correct for the case of a small cluster of matter. Simple but computationally intensive models of the motion of electrons and ions in such an illuminated cluster hint at more

complex, collective absorption mechanisms, but don't yet provide a full explanation. Wabnitz *et al.* have thus given fresh impetus for theorists to revisit the question of absorption of light by small particles — a question that is also relevant for areas such as nanotechnology, where light absorption by nanoscale systems is likely to be an important practical consideration. Physics with the new generation of FELs has begun. ■

Henry C. Kapteyn is in the Department of Physics and JILA, University of Colorado, Boulder,

Colorado 80309-440, USA.

e-mail: kapteyn@jila.colorado.edu

Todd Ditmire is in the Department of Physics, The University of Texas at Austin, 1 University Station C1600, Austin, Texas 78712, USA.

e-mail: tditmire@physics.utexas.edu

1. Wabnitz, H. *et al.* *Nature* **420**, 482–485 (2002).
2. Umstadter, D. *Phys. Plasmas* **8**, 1774–1785 (2001).
3. Kodama, R. *et al.* *Nature* **412**, 798–802 (2001).
4. Rundquist, A. *et al.* *Science* **280**, 1412–1415 (1998).
5. Pellegrini, C. *Nucl. Instrum. Methods A* **475**, 1–12 (2001).
6. Neutze, R. *et al.* *Nature* **406**, 752–757 (2000).
7. Ditmire, T. *et al.* *Nature* **398**, 489–492 (1999).

Immunology

Education and promiscuity

William R. Heath and Hamish S. Scott

Immune cells must be taught to distinguish between invading microbes and the body's own proteins. A new study re-emphasizes the importance of a thorough education in the thymus, and identifies an essential instructor.

To fight off microbial infections, our bodies call into action a type of immune cell known as the T cell. Sometimes, however, these cells mistake part of our body (self) for a microbe (non-self), and the resulting 'friendly fire' can lead to autoimmune diseases such as multiple sclerosis and juvenile diabetes. To prevent this, T cells must undergo a strict education in the thymus, where they are tested for reactivity to as many self-proteins as possible; if reactive, they are eliminated.

But how does this education process work? Writing in *Science*, Anderson and colleagues¹ report that organ-specific self-proteins (such

as insulin, from the pancreas) can be 'promiscuously' expressed in the thymus, under the control of the Aire protein. This allows developing T cells to become thoroughly familiar with many of the body's molecules, including those normally concealed in distant organs, and thus minimizes autoimmunity. These findings explain why humans with a non-functional form of this protein suffer a range of autoimmune diseases^{2,3}.

The story began more than a decade ago, with the identification of a rare recessive genetic disease⁴ — more common in some genetically isolated populations such as the Finnish, Sardinians and the Iranian Jews

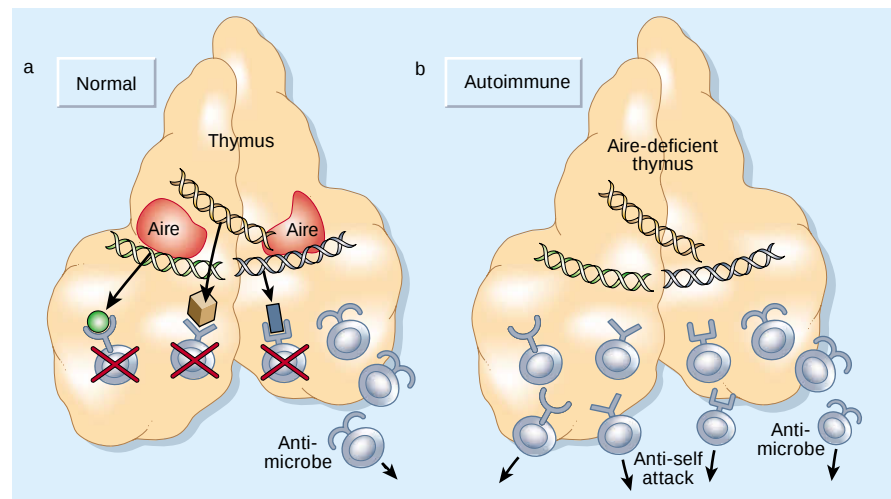


Figure 1 Aire and autoimmunity. Developing T cells must be educated in the thymus so that those cells that attack microbes survive, but cells that target self-proteins are destroyed. **a**, From their studies of mice lacking the Aire protein, Anderson *et al.*¹ propose that, in the thymus, this protein (AIRE in humans) drives the expression of proteins (represented by different shapes) that are found in distant organs. This casts a 'self shadow' on the thymus that is crucial in educating developing T cells: microbe-specific T cells survive, leave the thymus and fight infection, while self-specific T cells are eliminated. **b**, People with the genetic disorder APECED (who do not produce functional AIRE) and Aire-deficient mice do not express some organ-specific proteins in the thymus, allowing self-reactive T cells to escape into the body, where they eventually attack the organs in which their target proteins occur.

— that goes by the name of autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED), or autoimmune polyglandular syndrome type 1. The disorder is partially characterized by a tissue-specific destructive autoimmune process that principally affects the small hormone-secreting organs⁴, including the pancreas, which is responsible for insulin production. Unlike most autoimmune diseases, APECED is caused by a defect (mutation) in a single gene. As such, it was recognized that identification of this gene might shed light on how self-directed immune responses are regulated.

The gene mutated in APECED, termed the 'autoimmune regulator', was independently identified by two collaborations^{2,3} in 1997, and the gene product (AIRE in humans; Aire in mice) was later shown to control gene transcription *in vitro*^{5,6}. Although its expression pattern is not fully defined, AIRE has been reported to be produced by cells involved in educating and selecting T cells, such as thymic medullary epithelial cells and dendritic cells⁷. This pattern, together with the strong immunological basis for APECED, hinted at an underlying immune function for AIRE. Using a mouse model, Anderson *et al.*¹ now identify such a function for this transcription factor.

T cells recognize microbes by using specific T-cell receptor proteins. Although each T cell expresses just one type of T-cell receptor, these proteins vary greatly within a cell population, ensuring that all potential microbial infections can be detected. Because the structure of this receptor is generated at random as T cells mature in the thymus, some cells arise with receptors that react to self-proteins, instead of microbes. In most cases, such self-reactive T cells are eliminated by a process called negative selection, which occurs as part of the strict education in the thymus. For negative selection to work effectively, however, developing T cells must encounter self-proteins in the thymus. This raises the question of how self-reactive T cells that are specific for proteins occurring in organs far from the thymus, such as the pancreas or brain, can be eliminated.

One way in which this has been postulated to occur is through the promiscuous (ectopic) expression of organ-specific genes within the thymus^{8,9}. Anderson *et al.*¹ now make the crucial observation that the Aire protein regulates promiscuous thymic expression of organ-specific genes in mice (Fig. 1). The authors show that thymic epithelial cells from mice lacking functional Aire fail to promiscuously express many of the organ-specific genes normally found in the thymus, and are therefore unable to educate developing T cells properly. This allows self-reactive T cells to avoid thymic elimination, enter the peripheral immune system, and eventually cause autoimmune damage. These observations may explain

why APECED patients, who have mutant AIRE, show a range of autoimmune conditions. Interestingly, the array of tissues damaged by autoimmunity in humans and mice lacking this transcription factor is somewhat different, hinting at a contribution by other genetic or environmental factors.

This fascinating study¹ strongly implies that Aire-driven promiscuous expression of many organ-specific genes in the thymus is essential for limiting autoimmunity. It also raises many questions. First, Anderson *et al.* found that not all promiscuously expressed genes in thymic epithelial cells were extinguished in Aire-deficient mice. This suggests that other transcription factors may fulfil a similar role, although none with a similar combination of structural domains seems to be encoded in the human or mouse genomes. Furthermore, it is not clear how Aire regulates the expression of so many different genes in the thymus.

Second, we have so far discussed thymic education only in terms of the elimination of self-reactive T cells. But there is growing evidence that regulatory T cells that act to dampen immune responses are also generated in the thymus. Might Aire-induced thymic expression of organ-specific proteins also be necessary to promote the survival of regulatory T cells that dampen autoimmunity?

Condensed-matter physics

Rabi flopping sees the light

Thomas Udem

Ultrashort laser pulses are a valuable tool, but at these timescales a factor called the 'carrier-envelope phase' becomes important. A new technique to measure the evolution of this phase could advance laser spectroscopy.

Inside the envelope of a laser pulse is the rapidly cycling electric field of the carrier wave. If the laser pulse is very short — of the order of femtoseconds ($1 \text{ fs} = 10^{-15} \text{ s}$), as it is in many modern applications — the carrier wave consists of very few field cycles. In this situation, the relative positions of wave peaks and troughs (the 'phase'), between the envelope and carrier wave, come into play (Fig. 1). Surprisingly, although the absolute value of this carrier-envelope phase cannot yet be determined, it is possible to measure the change (or advance) of the phase between the pulses. If the phase advance can be controlled, this has implications for the precision of measurements in laser spectroscopy, as well as for the discovery of new phenomena. In *Physical Review Letters*, Oliver Mücke and colleagues¹ propose a new technique for measuring this phase advance that exploits the process of optical 'Rabi flopping'.

In 1936, Isidor Rabi published a groundbreaking study² of a two-level, quantum-

mechanical system. These levels might be two energy levels of an atom, the two possible orientations of a spin in a magnetic field, or the valence and conduction bands of a semiconductor. Rabi looked at how the occupation of the levels was affected when the system was stimulated by an electromagnetic wave of the same frequency as the natural transition frequency between the levels (Ω_0). He found that, when the electromagnetic wave was turned on, starting from the lowest-energy (ground) state the system became progressively more excited, with increasing occupation of the higher level. But once the system had become fully excited, the same electromagnetic wave stimulated the emission of radiation and the system returned to the ground state — the process then started all over again. This is Rabi flopping³, and it has its own frequency, Ω_R , which depends on the intensity of the electromagnetic wave.

More than 30 years later, Ben Mollow realized⁴ that the electromagnetic wave driving Rabi flopping in a two-level system is

Third, studies of another Aire-deficient mouse model suggested a role for Aire in self-tolerance mechanisms that take place in the peripheral immune system (once T cells leave the thymus)¹⁰. In this context, dendritic cells are known to express Aire and to be involved in both the thymic and peripheral education of T cells — but they were not found to contribute to self-tolerance in the study by Anderson *et al.*¹. Even so, is it possible that these cells make more subtle contributions? And finally, will other, more common autoimmune diseases show related defects in thymic education?

William R. Heath is in the Division of Immunology and Hamish S. Scott is in the Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia.

e-mails: heath@wehi.edu.au

hscott@wehi.edu.au

1. Anderson, M. S. *et al. Science* **298**, 1395–1401 (2002).

2. Nagamine, K. *et al. Nature Genet.* **17**, 393–398 (1997).

3. The Finnish–German APECED Consortium *Nature Genet.* **17**, 399–403 (1997).

4. Ahonen, P. *Clin. Genet.* **27**, 535–542 (1985).

5. Björns, P. *et al. Am. J. Hum. Genet.* **66**, 378–392 (2000).

6. Kumar, P. G. *et al. J. Biol. Chem.* **276**, 41357–41364 (2001).

7. Heino, M. *et al. Biochem. Biophys. Res. Commun.* **257**, 821–825 (1999).

8. Derbinski, J., Schulte, A., Kyewski, B. & Klein, L. *Nature Immunol.* **2**, 1032–1039 (2001).

9. Hanahan, D. *Curr. Opin. Immunol.* **10**, 656–662 (1998).

10. Ramsey, C. *et al. Hum. Mol. Genet.* **11**, 397–409 (2002).

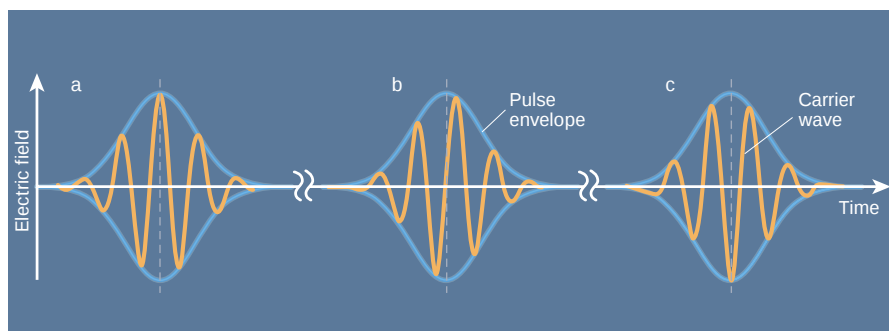


Figure 1 Phase slip. In ultrashort laser pulses (typically of femtosecond duration), the phase between the electric-field carrier wave and the pulse envelope becomes an important factor in experiments. The phase slips between pulses: for example, in a, the phase is zero; at a later time (b), the phase has slipped to 90° ; and later still (c), the phase is 180° . Mücke *et al.*¹ propose that this phase advance can be measured through optical Rabi flopping.

itself modified by the process. For example, if a laser with resonance frequency Ω_0 is used, the flopping between the levels modulates the amplitude of the scattered light, at the Rabi frequency, Ω_R . As a result, the spectrum of the scattered laser light no longer matches that of the laser pulse used to excite the system. As Mollow showed⁴, the outcome is two extra frequencies, called sidebands, which are separated from the original frequency of the incident wave by the Rabi-flopping frequency: the so-called Mollow triplet adds the frequencies $\Omega_0 + \Omega_R$ and $\Omega_0 - \Omega_R$ to Ω_0 .

The latest developments in 'mode-locked' laser technology have allowed extremely short (femtosecond) pulses to be produced, with the laser power concentrated at high peak intensities. As the Rabi frequency increases with laser intensity, in this new regime Ω_R is comparable to the rapid oscillations of the laser itself, which are of optical frequency, and a phenomenon called carrier-wave Rabi flopping takes over: the Rabi cycle is now driven by the carrier wave of the laser, not by the pulse envelope. In earlier work, Mücke *et al.*⁵ reported this rapid Rabi flopping in the semiconductor gallium arsenide, and similar effects in zinc oxide⁶.

In carrier-wave Rabi flopping, other frequencies in addition to the Mollow triplet appear in the scattered wave; these frequencies are odd multiples (harmonics) of the original laser frequency. Each harmonic also has sidebands. Mücke *et al.*¹ have now considered what happens if Ω_R becomes as high as Ω_0 — that is, if Rabi flopping occurs at optical frequencies. In this case, the higher-frequency Rabi sideband at the laser frequency ($\Omega_0 + \Omega_R$) and the lower-frequency sideband of the third, and strongest, harmonic ($3\Omega_0 - \Omega_R$) overlap, giving rise to interference between the scattered components at around a frequency of $2\Omega_0$. Mücke *et al.* propose that, by measuring this interference (which is feasible⁷), it should be possible to derive the carrier-envelope phase advance between the pulses (and perhaps ultimately the phase itself).

A mode-locked laser generally emits a

train of pulses with a fluctuating carrier-envelope phase, and the possibility of detecting and controlling the motion of this phase offers the opportunity to produce a train of truly identical pulses. If the carrier-envelope phase is stabilized, each laser mode becomes coherently linked to the repetition rate of the pulses. If the pulse repetition rate is then referenced to an atomic clock, all the laser modes are also referenced to that clock and are available for high-precision laser spectroscopy. A stable carrier-envelope phase would also be expected to reveal new phenomena in nonlinear optics^{8,9}.

The existing technique for measuring the carrier-envelope phase advance (though not the absolute phase value) relies on broadening the spectrum of the laser pulses in specially designed optical fibres. The lower-frequency part of the spectrum can then be 'frequency-doubled' and made to interfere with the higher-frequency part. The frequency-doubling also doubles the carrier-envelope phase advance, which can similarly be measured from the interference. This has become a standard technique, but the optical fibre itself is delicate and contributes to some instabilities, so researchers have been looking for more robust and simpler detection methods.

Other approaches have been devised, based on lasers^{10,11} that already put out a sufficiently wide range of frequencies to avoid the need for spectral broadening in a fibre. But such lasers can be difficult to handle and are not yet readily available. Alternatively, improved fibre technology could simplify the spectral-broadening system. Nevertheless, the proposal by Mücke *et al.*¹ to access the carrier-envelope phase advance through optical Rabi flopping is a welcome new avenue of exploration.

Thomas Udem is at the Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, 85748 Garching, Germany.
e-mail: thomas.udem@mpq.mpg.de

- Mücke, O. D., Tritschler, T., Wegener, M., Morgner, U. & Kärtner, F. X. *Phys. Rev. Lett.* **89**, 127401 (2002).
- Rabi, I. I. *Phys. Rev.* **49**, 324–328 (1936).



100 YEARS AGO

In your report of the meeting of the Physical Society of October 31, I find the following sentence given as having been said by me in the course of some remarks on Mr. Ridout's paper on the size of atoms, with the four words which I underline accidentally omitted. "If the electrons, or atoms of electricity, succeeded in getting out of the atoms of matter, they proceeded with velocities which might exceed the velocity of light, and the body was radioactive." The omission of those four words made it appear that I had considered the velocity of the escaping electrons to be essentially the velocity of light. In reality, the electrons may escape with velocities possibly less or possibly more than the velocity of light, but certainly not all with one definite velocity...

Kelvin

[The official report of Lord Kelvin's remarks was printed as received.— Editor.]
From *Nature* 4 December 1902.

50 YEARS AGO

The Darwin Medal is awarded to Prof. John Burdon Sanderson Haldane, Weldon professor of biometry in University College, London, for his work on the analysis of the causes of variation and of the mechanism of selection. The conclusions derived from his researches have permeated practically every field of evolutionary discussion, and his ideas have fundamentally altered our knowledge of evolutionary change. For many years he has insisted that a proper study of evolution involves the development of methods for investigation of the genetics of populations. His mathematical treatment of the question was the first of its kind, and has played a large part in the great development of understanding of evolution during the past twenty-five years. He has been a pioneer in applying biochemical knowledge... to problems of genetics. His contributions to strictly genetical problems are fundamental ... He made the first investigations of human mutation-rates (simultaneously with L. S. Penrose), has studied interspecific hybrids and discovered the rule by which their sex is determined. He was one of the first to develop a theoretical treatment of polyploidy and to show its evolutionary significance. He has thus made first-rate contributions by his detailed researches, his mathematical treatments and his general analysis of evolutionary problems, to the field of Darwin's work.

From *Nature* 6 December 1952.

3. Cohen-Tannoudji, C., Diu, B. & Laloë, F. *Quantum Mechanics* 412 (Wiley, New York, 1977).
4. Mollow, B. R. *Phys. Rev.* **188**, 1969–1975 (1969).
5. Mücke, O. D., Tritschler, T., Wegener, M., Morgner, U. & Kärtner, F. X. *Phys. Rev. Lett.* **87**, 57401 (2001).
6. Mücke, O. D., Tritschler, T., Wegener, M., Morgner, U. & Kärtner, F. X. *Opt. Lett.* **27**, 2127–2129 (2002).

7. Udem, T., Holzwarth, R. & Hänsch, T. W. *Nature* **416**, 233–237 (2002).
8. Brabec, T. & Krausz, F. *Rev. Mod. Phys.* **72**, 545–591 (2000).
9. Paulus, G. G. *et al.* *Nature* **414**, 182–184 (2001).
10. Ramond, T. M., Diddams, S. A., Hollberg, L. & Bartels, A. *Opt. Lett.* **27**, 1842–1844 (2002).
11. Morgner, U. *et al.* *Phys. Rev. Lett.* **86**, 5462–5465 (2001).

Conservation biology

Lone wolf to the rescue

Pär K. Ingvarsson

Genetic analysis has revealed how a small and isolated population of grey wolves found salvation in the form of the genetic variation offered by a single, immigrant male.

Human influence has led to the decline of many animal species and fragmentation of their populations. One consequence is loss of their genetic diversity. But how does this loss affect endangered species? Some have argued that it is only a minor threat to survival, because small populations are more vulnerable to demographic or environmental effects — for example, chance events affecting the birth and death of individuals, or atypical weather conditions such as an unusually cold winter. Others believe that a lack of genetic diversity increases the risk of inbreeding, which may ultimately drive a species to extinction because inbred organisms have reduced survival and fecundity.

Immigration from surrounding areas may prevent the extinction of small populations by increasing their size. Alternatively, genetic rescue — replenishing genetic variation and reducing the consequences of inbreeding — could have the same effect. A paper by Vilà *et al.*¹, published last month by *Proceedings of the Royal Society*, lends strong support to the view that genetic rescue does occur². The authors provide compelling evidence that the current Scandinavian population of the grey wolf (*Canis lupus*; Fig. 1) was for a long time limited in size by lack of genetic diversity. Furthermore, the authors

show that the steady increase in that population over the past decade started with the arrival of a single immigrant wolf.

Scandinavian grey wolves were decimated during the nineteenth and twentieth centuries, largely through hunting and poisoning. This population was finally considered to be extinct in the 1960s. But in the early 1980s, a single breeding pack was once again established in southern Scandinavia, about 1,000 km from the nearest known populations in Finland and Russia. The pack remained small, consisting of around ten animals, for several years. It was monitored from its initial establishment, and genetic data showed that a single male and female were the founding members. Moreover, the population suffered from inbreeding and declining levels of genetic variation over time. At the same time, studies of captive wolves indicated that birth defects, such as hereditary blindness, were common for levels of inbreeding similar to those observed in the wild population^{3,4}.

Suddenly, in 1991, the Scandinavian wolf population started to increase and now numbers 10–11 breeding packs, totalling about 100 wolves. From the genetic data obtained by Vilà and colleagues¹, it is evident that the increase corresponds with the immigration of a single male wolf of Finnish or

Siberian origin. Coinciding with the arrival of this male, much of the genetic diversity that had been lost from the population was restored. Vilà *et al.* show that of the 72 wolves born after 1993, 68 can trace at least part of their ancestry to the immigrant male of 1991. This male has had a disproportionately large influence on the genetic make-up of the present Scandinavian wolf population, implying that natural selection⁵ has driven the rapid increase in numbers.

The authors hint at a possible cause for the extraordinary reproductive success of the immigrant male. Many species have behavioural mechanisms that prevent matings between close relatives to avoid the deleterious effects of inbreeding, and wolves are no exception. So these mechanisms may have contributed to the low birth rate in the original breeding pack and thus prevented the population from expanding. With the input of new genetic variation from the immigrant, the potential for matings between unrelated animals was considerably increased.

The study by Vilà *et al.*¹ follows others^{6,7} in suggesting that low levels of migration between populations of endangered species (either natural, or directed through relocation programmes) can be extremely effective in restoring genetic diversity and reducing the risk of extinction through inbreeding. But a cautionary note is needed here. In relocation programmes, it is essential that the animals are not moved to environments to which they are not adapted. Such action could actually increase the risk of extinction through a process known as outbreeding depression⁸, where offspring between animals adapted to different conditions have lower survival rates and lower fecundity in both parental environments because they are not adapted to either of them. Animals for relocation should be chosen from environments that are as similar as possible to their new habitat.

The relative merits of genetics and ecology as factors in understanding the risks of extinction continue to be debated. The study by Vilà *et al.* is a refreshing reminder of how important it is to combine both perspectives toward understanding what is required to preserve Earth's biodiversity.

Pär K. Ingvarsson is in the Department of Ecology and Environmental Sciences, University of Umeå, SE-901 87 Umeå, Sweden.

e-mail: pelle@eg.umu.se

1. Vilà, C. *et al.* *Proc. R. Soc. Lond. B* **10.1098/rspb.2002.2184** (2002).
2. Ingvarsson, P. K. *Trends Ecol. Evol.* **16**, 62–63 (2001).
3. Ellegren, H. *Hereditas* **130**, 239–244 (1999).
4. Laikre, L. & Ryman, N. *Conserv. Biol.* **5**, 33–40 (1991).
5. Ingvarsson, P. K. & Whitlock, M. C. *Proc. R. Soc. Lond. B* **267**, 1321–1326 (2000).
6. Madsen, T., Shine, R., Olsson, M. & Wittzell, H. *Nature* **402**, 34–35 (1999).
7. Westemeier, R. L. *et al.* *Science* **282**, 1695–1698 (1998).
8. Shields, W. M. in *The Natural History of Inbreeding and Outbreeding* (ed. Thornhill, N. W.) 143–169 (Univ. Chicago Press, 1993).



Figure 1 A Scandinavian wolf in the wild. Additional genetic variety has turned out to be the spice of life for the population concerned.

Hydrodynamics

Bend and survive

Victor Steinberg

The aim of aerodynamic design is to reduce the drag experienced by a body, such as a car, in a flowing medium, such as air. But what happens if the body is flexible and bends in response to the flow?

Everyone has seen trees, particularly young ones, bend in a strong wind. But fewer people realize that this is a clever way of reducing the drag force of the wind on the tree. Thus trees combine a functional necessity to grow tall and carry out photosynthesis with the ability to survive strong winds (Fig. 1). But just how much can a flexible object reduce drag by changing its shape? On page 479 of this issue, Alben, Shelley and Zhang¹ present results from their experimental and theoretical studies of drag reduction by shape reconfiguration for a flexible body immersed in a flowing medium.

When the velocity of a flowing medium increases, there is always an accompanying increase in the resistance of a body to the flow². The effect is even greater if the flow velocity exceeds some critical value and becomes turbulent, which is often the case in natural phenomena and engineering applications. The most effective way to counteract flow resistance is to design a body whose shape minimizes drag by reducing, or delaying, the turbulence in the boundary layer of the flow nearest the moving body. This is reflected in the design of aerodynamic shapes — from planes and trains, to cars and cyclists' helmets.

But these examples are all bodies with a fixed shape. A more ingenious way to reduce

drag, as seen in nature — from trees to jellyfish to amoebae — is for an immersed body to adjust its shape as the flow velocity increases. The only engineering application of this effect that immediately comes to mind is a sail — although, in this case, the reshaping of the sail in a stiff wind actually increases the drag, and hence drives the boat.

The outstanding feature of the studies presented by Alben *et al.*¹ is that they have designed an experiment that can be closely described by a theoretical model. The authors studied the behaviour of a two-dimensional flexible fibre immersed in flowing soapy water. The soap film makes the forces on the fibre easily visible (see Fig. 2 on page 480), and the flow velocity and the force on the fibre can be measured simultaneously. Using a model that couples the flow dynamics with the bending of the fibre, Alben *et al.* have derived a differential equation describing the curvature of the fibre in the flow that can be solved for any given value of a control parameter. This parameter is the non-dimensional velocity, η , defined as the ratio of the fluid kinetic energy to the bending elastic energy of a flexible body.

The model predicts a sharp transition from a nearly straight fibre to a bending one when the flow kinetic energy becomes comparable to the fibre's elastic bending energy — that is, when the value of the control parameter is of order one. One of the consequences of this transition is a change in the functional dependence of the drag on η , an effect that Alben *et al.* also see experimentally.

The particularly striking feature of the transition shows up in the shape adopted by the bending fibre. The fibre takes on a quasi-parabolic shape, and, if the overall length of the fibre (whatever its value) is scaled by a factor $\eta^{2/3}$, the parabola is always the same (Fig. 2). As η increases, so the increased bending of the fibre is confined to a progressively smaller region at its ends; the size of this region is related to a new parameter in the problem, the 'bending length' of the fibre. This length parameter arises from the competition between the rigidity of the fibre and the force of the flow and sets the 'self-similarity' scale for the parabolic bending shape. For large values of η , the total drag on the fibre is created by the tip region defined by the bending length, and not by the reduced fibre-profile width as might be assumed.

There are limitations in the comparison

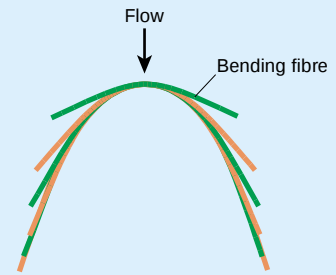


Figure 2 The curvature of a flexible fibre in a flowing medium increases with the strength of the flow. Alben *et al.*¹ show that, in fact, the curvature increases only around the tips of the fibre, and the central region maintains a quasi-parabolic, self-similar shape.

between the experiment and the model, some of which Alben *et al.* acknowledge in their paper. For example, the compressibility of the soap film, a result of thickness variations and skin friction, can alter the scaling behaviour. Another factor that is not mentioned is the surface tension between the soap film and the fibre. These effects probably account for the almost uniform systematic factor that Alben *et al.* have to apply to achieve a good quantitative comparison between the theory and their data.

There is certainly more to be learned from further studies of fibre bending, such as whether hysteresis (a difference between bending and unbending) occurs. It would also be interesting to extend the experiment to three dimensions, using a flexible body made of a gel with an appropriate elastic modulus, immersed in a three-dimensional flow. Whether the transition to bending and the self-similar shape survive in higher dimensions could then be verified.

Victor Steinberg is in the Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel.

e-mail: victor.steinberg@weizmann.ac.il

1. Alben, S., Shelley, M. & Zhang, J. *Nature* **420**, 479–481 (2002).
2. Landau, L. D. & Lifshitz, E. M. *Fluid Mechanics* (Pergamon, Oxford, 1987).

correction

In Hans-Georg Rammensee's article "Immunology: Survival of the fitters" (*Nature* **419**, 443–445; 2002), it was said that MHC class I ligands (peptide fragments that convey information to the immune system) are made up of exactly nine amino acids. This was a simplification, in that all copies of a particular MHC class I ligand are of a uniform length, and most are exactly nine amino acids long. Some, however, are composed of eight, ten or eleven amino acids, and there may be other exceptions.

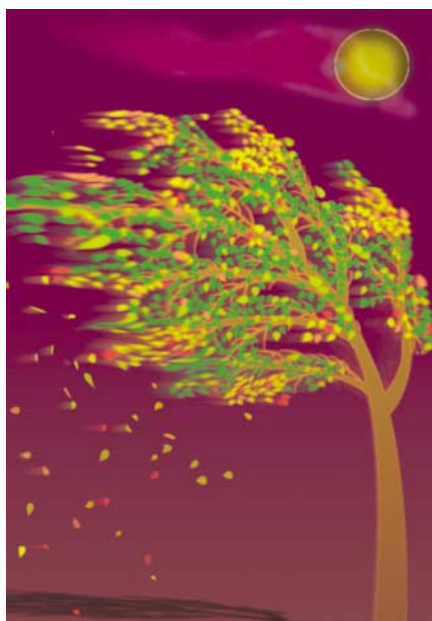


Figure 1 A tree bends in a gusty wind, reducing the potentially damaging drag force of the moving air on its branches.

Tree-hole frogs exploit resonance effects

These anurans know a trick or two when it comes to the romantic powers of song.

Animal mating calls that exert a comparatively high sound pressure propagate over greater distances and generally have greater attractive power^{1,2}. Here we show that calling male Bornean tree-hole frogs (*Metaphrynella sundana*) actively exploit the acoustic properties of cavities in tree trunks that are partially filled with water and which are primarily used as egg-deposition sites. By tuning their vocalizations to the resonant frequency of the hole, which varies with the amount of water that it contains, these frogs enhance their chances of attracting females.

In a simulated tree-hole experiment, we placed a calling male in an opaque plastic tube that was partially filled with water. The frog's vocalizations were recorded the following night as the water level was slowly reduced (increasing the air column from 50 to 144 mm) over a period of 28 min. We measured the change in air-column depth from a graduated cylinder connected to the tube, and later sampled time segments from the recording at regular intervals.

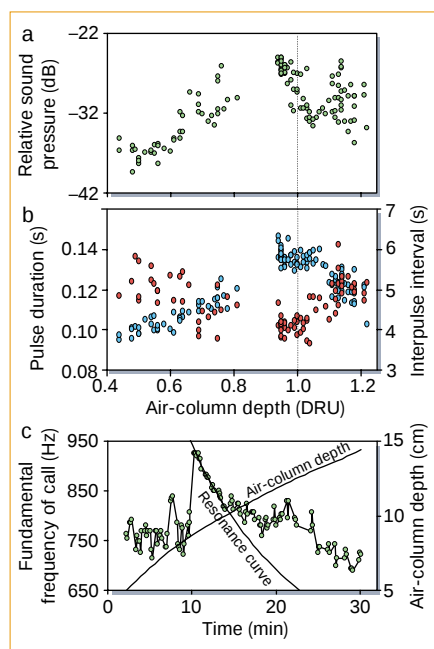


Figure 1 Resonance effect and associated temporal call and pitch alterations of a tree-hole frog calling from inside a tube as the air-column depth increases. **a**, A sound-pressure peak (resonance effect) is apparent at about 0.95 dynamic resonance units (DRU; the deviation from 1 DRU may be the result of experimental error in the levelling of the tube and graduated cylinder). **b**, At the resonance peak, the frog produced longer call pulses (blue) and the interpulse interval (red) was at a minimum. **c**, Call pitch variation over time. The resonance curve indicates the pitch values that give rise to resonance effects as the air-column depth increases (adjusted for the occurrence of resonance effects at 0.95 rather than 1.00 DRU).

The frog's call is a simple tonal pulse with little or no frequency modulation; the fundamental frequency is the dominant one. We analysed the call's pitch in relation to the air-column depth in dynamic resonance units (DRU), calculated as the call pitch at any point divided by the resonant frequency at that point. Resonance was predicted to occur at 1 DRU (corresponding to an air-column depth of one-quarter of the wavelength of the current pitch).

A peak in sound pressure was evident at about 0.95 DRU (Fig. 1a). The pulse duration peaked when the resonance effect was strongest (Fig. 1b). Inspection of oscillograms indicated that this was not merely due to an echo effect but to the production of longer call pulses (results not shown). Also, the interval between call pulses was shortest when the sound pressure peaked (Fig. 1b). Both of these call alterations are energetically costly³ and are known to increase male attractiveness in many amphibians^{4–6}.

In addition to its flexibility in calling effort, the frog initially called with a varying pitch until it found the frequency that resulted in maximum amplitude. It then managed to track closely the falling water level in its resonance-matching over a period of several minutes, gradually lowering its pitch by about 115 Hz (Fig. 1c). The frog then apparently lost track of the optimal call pitch and resumed calling with an erratic pitch. Although this result relates to only one individual, the most likely explanation for this striking pattern is that the frog was actively tracking the resonant frequency of the tube.

Our field recordings of males calling from natural tree holes (Fig. 2) also show this correlation between pulse duration and interpulse interval, and between pulse duration and call pitch variability (Pearson correlation coefficient of -0.245 , $n = 238$, $P < 0.001$; and correlation coefficient of -0.237 , $n = 237$, $P < 0.001$, respectively; pitch variability is given as pitch coefficient of variance based on 10 ± 4 (mean $\pm$ s.d.) consecutive call pulses). This indicates that frogs in natural tree holes also increase their calling effort if they can adjust their call pitch to match the resonant properties of the hole.

Several crickets^{7–9} and burrowing frogs^{10,11} benefit from sound amplification by calling from baffles or burrows. However, to our knowledge this is the first evidence of an animal not only sampling resonance properties¹² but also facultatively adjusting its call pitch and calling strategy



Figure 2 Lover's lair: the tree-hole frog uses acoustic acumen to ensure that his mating calls strike a chord with local females.

in what seems to be an adaptive manner. Our findings indicate that animals may be better at exploiting signal-enhancing structures than was previously appreciated, and that flexibility in related, but not obligately linked, traits may also follow from such adaptive strategies.

Björn Lardner*, Maklarin bin Lakim†

*Division of Amphibians and Reptiles, Field Museum of Natural History, Chicago, Illinois 60605, USA

e-mail: bjorn.lardner@zoookol.lu.se

†Research and Education Division, Sabah Parks, PO Box 10626, 88806 Kota Kinabalu, Sabah, Malaysia

- Gerhardt, H. C., Dyson, M. L. & Tanner, S. D. *Behav. Ecol.* **7**, 7–18 (1996).
- Gerhardt, H. C., Tanner, S. D., Corrigan, C. M. & Walton, H. C. *Behav. Ecol.* **11**, 663–669 (2000).
- Wells, K. D. in *Anuran Communication* (ed. Ryan, M. J.) 45–60 (Smithsonian Inst. Press, Washington DC, 2001).
- Welch, A. M., Semlitsch, R. D. & Gerhardt, H. C. *Science* **280**, 1928–1930 (1998).
- Klump, G. M. & Gerhardt, H. C. *Nature* **326**, 286–288 (1987).
- Gerhardt, H. C. & Huber, F. *Acoustic Communication in Insects and Anurans* (Univ. Chicago Press, Chicago, 2002).
- Bennet-Clark, H. C. *J. Exp. Biol.* **128**, 383–409 (1987).
- Bailey, W. J., Bennet-Clark, H. C. & Fletcher, N. H. *J. Exp. Biol.* **204**, 2827–2841 (2001).
- Prozesky-Schulze, L., Prozesky, O. P. M., Anderson, F. & van der Merwe, G. J. *Nature* **255**, 142–143 (1975).
- Bailey, W. J. & Roberts, J. D. *J. Nat. Hist.* **15**, 693–702 (1981).
- Penna, M. & Solis, R. *Anim. Behav.* **51**, 255–263 (1996).
- Daws, A. G., Bennet-Clark, H. C. & Fletcher, N. H. *Biocoustics* **7**, 81–117 (1996).

Competing financial interests: declared none.

Mathematics

What is the best way to lace your shoes?

The two most popular ways to lace shoes have historically been to use 'criss-cross' or 'straight' lacing — but are these the most efficient? Here we demonstrate mathematically that the shortest lacing is neither of these, but instead is a rarely used and unexpected type of lacing known as 'bowtie' lacing. However, the traditional favourite lacings are still the strongest.

The $2n$ eyelets of an idealized shoe are the points of intersection of two vertical lines and n equally spaced horizontal lines in the plane. The two columns of eyelets are one unit apart, and two adjacent rows of eyelets are a distance h apart. An n -lacing of our shoe is a closed path in the plane that consists of $2n$ line segments, the end points of which are the $2n$ eyelets.

For any given eyelet, we require that at least one of the two segments that ends in it should not be contained in the same column as that eyelet; this condition ensures that every eyelet genuinely contributes towards pulling the two sides of the shoe together. Virtually all lacings that are actually used satisfy this condition.

We call a lacing 'dense' if neither of the

two segments ending in any eyelet is contained in the same column as the eyelet — that is, a dense lacing zigzags back and forth between the two columns of eyelets as, for example, do the traditional lacings (Fig. 1a, b). Finally, we assume that n is at least 2.

Using standard combinatorial techniques, we find that the number of n -lacings is

$$\frac{(n!)^2}{2} \sum_{k=0}^m \frac{1}{n-k} \binom{n-k}{k}$$

where $m = n/2$ for even n , and $m = (n-1)/2$ for odd n . The number of dense n -lacings is

$$\frac{n!(n-1)!}{2}$$

The length of an n -lacing is the sum of the lengths of the segments that it consists of. Using the symmetries of the configuration of eyelets, it is possible to design a powerful list of local shortening rules and to use these to identify the bow-tie n -lacings as the shortest n -lacings (Fig. 1c–e). Furthermore, by generalizing earlier results^{1–4}, we can show that the criss-cross n -lacing is the shortest dense n -lacing, even if the eyelets are not fully aligned. Note that it is also possible to identify the longest dense n -lacings for general n .

When you pull on the ends of a shoelace, it acts like a pulley. Ideally, the tension along the shoelace is a positive constant, T . This tension gives rise to a total tension, T_{hor} , of the pulley in the horizontal direction; that is, the direction in which the two sides of the shoe are being pulled together. This total tension, T_{hor} , is the sum of all horizontal components of T along the different segments of the lacing. The strongest n -lacings are then n -lacings that maximize T_{hor} .

The unique dense 2-lacing is also the strongest 2-lacing. Note that the shortest n -lacing is independent of the distance h between two adjacent rows of eyelets. In contrast, for $n > 2$, the strongest n -lacing does depend on h . We can show that there is a positive value, h_n , such that the strongest n -lacings are: the criss-cross n -lacing, for $h < h_n$; the criss-cross n -lacing and the straight n -lacings, for $h = h_n$; and the straight n -lacings, for $h > h_n$.

For many real shoes with n pairs of eyelets, the ratio of the distance between adjacent rows of eyelets and the distance between the columns is very close to h_n . This means that no matter whether you prefer to lace them straight or criss-crossed, you come close to maximizing the total horizontal tension when you pull on the two ends of one of your shoelaces.

And what is the strongest way to tie your shoelaces?⁵ Most people place one half-granny knot on top of another (it is not

essential to consider the loops here), which results in either a notoriously unstable granny knot or a very stable reef knot, depending on whether the two half-knots have the same or opposite orientation. As we have seen, hundreds of years of trial and error have led to the strongest way of lacing our shoes, but unfortunately the same cannot be said about the way in which most of us tie our shoelaces — with a granny knot.

Burkard Polster

School of Mathematical Sciences, PO Box 28M,
Monash University, Victoria 3800, Australia
e-mail: burkard.polster@sci.monash.edu.au

1. Halton, J. H. *Math. Intell.* **17**, 37–41 (1995).
2. Misiurewicz, M. *Math. Intell.* **18**, 32–34 (1996).
3. Stewart, I. *Sci. Am.* 78–80 (July 1996).
4. Stewart, I. *Sci. Am.* 86 (December 1996).
5. Smith, G. A. www.u.arizona.edu/~gasmith/knots/knots.html

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Laser-Raman spectroscopy

Images of the Earth's earliest fossils?

Fossil remains of the most ancient, minute forms of life on Earth and other planets are hard to recognize. Schopf *et al.*¹ claim to have identified the biological remnant material known as kerogen in microscopic entities in rock by using Raman spectroscopic analysis. On the basis of a substantial body of published evidence, however, we contend that the Raman spectra of Schopf *et al.*¹ indicate that these are disordered carbonaceous materials of indeterminate origin. We maintain that Raman spectroscopy cannot be used to identify microfossils unambiguously, although it is a useful technique for pinpointing promising microscopic entities for further investigation.

We believe that Schopf *et al.* have over-interpreted their Raman spectra in attributing biogenicity to the extremely ancient, fossil-like objects that they analysed, as already addressed by Brasier *et al.*² We disagree with the underlying assumption by Schopf *et al.* that Raman spectroscopy is sensitive to a distinctive carbon signal of organic matter (kerogen), and with their conclusion that "measurement of Raman point spectra (Fig. 3) [together with optical microscopy and Raman imaging] substantiates the biological origin of the oldest putative fossils".

Contrary to the inference by Schopf *et al.*¹, laser-Raman microprobe spectroscopy does not reveal the chemical composition (as defined by geochemists) of a sample, but rather provides information about the molecular bonds of the constituent structural units. For instance, it can distinguish between carbon bonds in carbonate, in condensed benzene rings, in graphite and

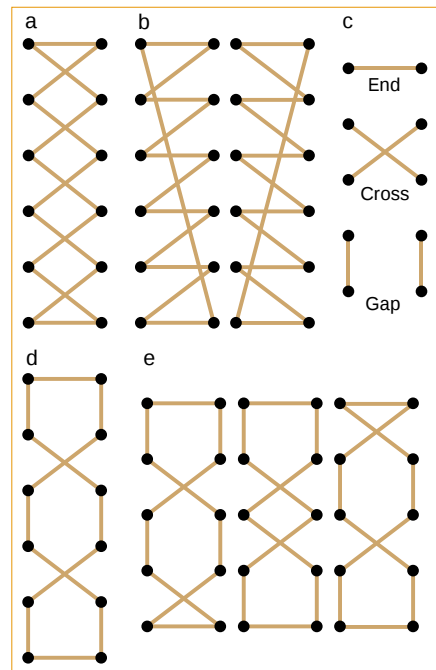


Figure 1 The strongest and the shortest shoe lacings. **a, b**, The most popular n -lacings, the criss-cross n -lacing (**a**) and the two straight n -lacings (**b**) are also the strongest n -lacings (here n is the number of pairs of eyelets). **c–e**, The shortest n -lacings are the bow-tie n -lacings. They are made up of ends, gaps and crosses (**c**). If n is even, there is exactly one bow-tie n -lacing (**d**) consisting of the two ends at the top and bottom, $n/2$ gaps and $n/2 - 1$ crosses. If n is odd, there are exactly $(n+1)/2$ different bow-tie n -lacings (**e**) consisting of the two ends, $(n-1)/2$ gaps and $(n-1)/2$ crosses.

in diamond. It also characterizes the degree of crystalline order in solids. However, it does not provide a list of elemental components or even ratios of carbon, hydrogen and oxygen, which might help to establish the biogenicity of a compound.

Most pertinent to the analysis of putative microfossils is the fact that the Raman spectrum of carbonaceous (that is, carbon-dominated) materials is sensitive to the degree of ordering of the carbon they contain (distinguishing, for example, between amorphous carbon, poorly ordered graphitic material and highly crystalline graphite)^{3–10}. The Raman spectra of Schopf *et al.*¹ confirm that their samples consist of highly disordered carbonaceous material and are consistent with the spectra of kerogens^{3,4}. However, their spectra are indistinguishable from those of many other types of structurally disordered carbonaceous matter generated from a wide range of starting materials by a wide variety of processes. Those processes (including high-temperature heating of organic or inorganic compounds^{4–7}, inorganic deposition from high-temperature synthetic fluids^{3–5} and geological deposition from hydrothermal solutions⁸) and materials (for example, ion-bombarded graphite⁹ and graphite-intercalation compounds¹⁰) may be strictly non-biogenic. There are no distinctive features in the spectra shown by Schopf *et al.*¹ that directly and unambiguously link them to kerogens.

Raman microprobe spectroscopy is useful for investigating the molecular structure of micrometre-sized features, such as putative microfossils, in rock. Showing that fossil-like objects consist of highly disordered carbonaceous material by Raman spectroscopy provides necessary, but not sufficient, evidence that the objects are biogenic. Although the microscopic objects analysed by Schopf *et al.*¹ may indeed be biogenic, we see nothing in their spectra that indicates the origin of their disordered carbonaceous material. The basic question remains unanswered: which measurable chemical and/or physical properties of a fossilized and/or altered material will unambiguously identify it as biological in origin?

Jill Dill Pasteris, Brigitte Wopenka

Department of Earth and Planetary Sciences,
Washington University, St Louis,
Missouri 63130-4899, USA
e-mail: pasteris@levee.wustl.edu

Shanks, W. C. & Ridley, W. I. 233–250 (Soc. Econ. Geol., Littleton, Colorado, 1998).

9. Elman, B. S., Shayegan, M., Dresselhaus, M. S., Mazurek, H. & Dresselhaus, G. *Phys. Rev. B* **25**, 4142–4156 (1982).
10. Dresselhaus, M. S. & Dresselhaus, G. in *Light Scattering in Solids III* 2–57 (Springer, New York, 1982).

Schopf *et al.* reply — The criticism by Pasteris and Wopenka of our use of laser-Raman imagery to investigate the carbonaceous make-up of extremely ancient fossils¹ focuses only on their Raman signature; however, our interpretation that the carbonaceous matter that makes up these specimens is biogenic is based on several lines of evidence, of which Raman spectroscopy is only one.

We did not state, nor did we imply, that Raman spectroscopic analysis can by itself be used to establish the biological origin of geochemically highly altered carbonaceous matter present in ancient sediments. We believe that the biogenicity of such matter, whether in fossil-like objects or sapropel-like detritus, should be demonstrated by a combination of data drawn from independent but mutually reinforcing lines of evidence.

For fossils in each of the four geological units we analysed¹ — including those of the roughly 3,375-million-year (Myr)-old Kromberg Formation and 3,465-Myr-old Apex Chert, which are among the oldest fossils known — three lines of evidence are most compelling. These are their cellular morphology^{2,3}, their carbonaceous molecular-structural make-up^{1–4}, and the carbon isotope composition of such fossils⁵ and/or of co-existing particulate kerogen^{6,7}, which have been shown by replicate analyses^{5–7} to be well within the range established for Precambrian biological organic matter on the basis of over 1,200 measurements from hundreds of fossil-bearing units⁷.

Our study¹, which focuses on the first two of these lines of evidence, is centred on the use of laser-Raman imagery (rather than on more conventional single-point measurements), a technique new to palaeobiology^{1,4}. We showed that there is a one-to-one correlation of cellular morphology and carbonaceous make-up in individual microscopic fossils from each of the four units investigated. Our claim is that such an analysis based on a combination of morphology and chemistry together provides a powerful new means to investigate the biogenicity of putative fossil-like objects, a problem that for many decades has plagued the search for evidence of early life⁸.

J. William Schopf*, Anatoliy B. Kudryavtsev†, David G. Agresti‡, Thomas J. Wdowiak‡, Andrew D. Czaja*

*Department of Earth & Space Sciences, and
Institute of Geophysics & Planetary Physics,
University of California, Los Angeles,
California 90095-1567, USA
e-mail: schopf@ess.ucla.edu

†Department of Physics, and ‡Astro and Solar
System Physics Program, Department of Physics,
University of Alabama, Birmingham,
Alabama 35294-1170, USA

1. Schopf, J. W., Kudryavtsev, A. B., Agresti, D. G., Wdowiak, T. J. & Czaja, A. D. *Nature* **416**, 73–76 (2002).
2. Schopf, J. W. in *The Proterozoic Biosphere, a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 25–39, 1055–1117 (Cambridge Univ. Press, New York, 1992).
3. Schopf, J. W. *Science* **260**, 640–646 (1993).
4. Kudryavtsev, A. B., Schopf, J. W., Agresti, D. G. & Wdowiak, T. J. *Proc. Natl Acad. Sci. USA* **416**, 73–76 (2002).
5. House, C. H. *et al. Geology* **28**, 707–710 (2000).
6. Strauss, H. & Moore, T. B. in *The Proterozoic Biosphere, a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 709–798 (Cambridge Univ. Press, New York, 1992).
7. Strauss, H., Des Marais, D. J., Hayes, J. M. & Summons, R. E. in *The Proterozoic Biosphere, a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 117–127 (Cambridge Univ. Press, New York, 1992).
8. Schopf, J. W. & Walter, M. R. in *Earth's Earliest Biosphere, its Origin and Evolution* (ed. Schopf, J. W.) 214–239 (Princeton Univ. Press, Princeton, New Jersey, 1983).

COMMUNICATIONS ARISING

Palaeontology

Thermal alteration of the Earth's oldest fossils

Microscopic carbonaceous structures found in ancient rocks could provide clues to early life on Earth if they turn out to be genuine fossil micro-organisms. Here we show that thermal alteration of microbial remains embedded in a mineral matrix may significantly change their original morphology and produce structures that resemble those of what are claimed to be the Earth's oldest fossils¹. These observations may shed light on the controversy^{2,3} that surrounds these microfossils from the 3,465-Myr-old Apex Chert of the early Archaean Warrawoona Group in northwestern Australia.

The biogenicity of these fossils has been called into question³ on the basis of suggestions that the Apex Chert structures were formed from amorphous graphite within multiple generations of metalliferous hydrothermal-vein chert and volcanic glass, and that the carbonaceous composition and characteristically biological carbon-isotope make-up of the carbonaceous filaments could have been products of non-biological (Fischer–Tropsch) organic synthesis³.

We have investigated structures that are present in silicified (chertified) cyanobacterial mats from the Bardo Mountains (Żdanów locality⁴) in southwestern Poland, which date to the early Silurian period (about 440 million years ago). The fossil mats occur in black, laminated radiolarian cherts, which have been interpreted as sediments that formed at moderate depths within the photic zone⁵. The mats are composed of cyanobacteria that are closely related to representatives of modern colonial chroococcaleans (particularly the

1. Schopf, J. W., Kudryavtsev, A. B., Agresti, D. G., Wdowiak, T. J. & Czaja, A. D. *Nature* **416**, 73–76 (2002).
2. Brasier, M. D. *et al. Nature* **416**, 76–81 (2002).
3. Wopenka, B. & Pasteris, J. D. *Am. Mineral.* **78**, 533–557 (1993).
4. Beny-Bassez, C. & Rouzaud, J. N. *Scan. Electr. Microsc.* 119–132 (1985).
5. Vidano, R. & Fischbach, D. B. *J. Am. Ceram. Soc.* **61**, 13–17 (1978).
6. Sato, Y., Kamo, M. & Setaka, N. *Carbon* **16**, 279–280 (1978).
7. Lespade, P., Al-Jishi, R. & Dresselhaus, M. S. *Carbon* **20**, 427–431 (1982).
8. Pasteris, J. D. in *Applications of Microanalytical Techniques to Understanding Mineralizing Processes* (eds McKibben, M. A.,

families Entophysalidaceae and Xenococcaceae⁶). Living colonies of these cyanobacteria are composed of globular subcolonies surrounded by thick mucous envelopes (Fig. 1a). The subcolonies are composed of minute cells (which in some species are less than 2 μm in diameter).

Post mortem degradation processes in modern mats composed of coccoid cyanobacteria⁷ indicate that the components that are most resistant to decay are the thicker outer mucous envelopes that surround groups of cells, subcolonies and entire colonies. After burial, these partially biodegraded envelopes often remain preserved as a cobweb-like polysaccharide material. With time and progressive diagenesis, this material may undergo kerogenization and be transformed from a more-or-less structured biological material into amorphous organic matter.

The early-Silurian cyanobacterial mats we describe represent a kerogenized stage in which the outlines of the subcolonies and of even smaller groups of densely packed, minute cells are still recognizable in the chert matrix (Fig. 1e). Owing to compaction, the circular outlines of the subcolonies are best seen in petrographic thin sections made parallel to the bedding. The subcolonies reach 60–90 μm in diameter and occur as blackish, cobweb-like structures (Fig. 1b; and see Fig. 1c in supplementary information) or as yellow–brown circular or semicircular areas, which are bordered by blackish, continuous or discontinuous, irregularly segmented or porous zones (Fig. 1c, d; and see Fig. 1a, b in supplementary information).

The densely packed masses of small bodies that fill the interiors of many subcolonies (Fig. 1e) are likely to be remnants of cells that created the original colonies. In vertical thin sections, the blackish material is usually present as slightly curved or undulated, often segmented, filamentous structures (Fig. 1f) which are locally branched (see Fig. 2 in supplementary information).

All of these features make the blackish structures almost identical in appearance to the filamentous structures described from the Apex Chert. This similarity is particularly striking when the shapes of Archaeal structures are compared to the porous and irregularly segmented structures from the peripheries of the Silurian cell aggregates. Their quasi-circular, C- and J-shaped outlines fit almost perfectly the morphologies described for the Apex structures^{1,3}. The same is true for the size classes of both groups of structures.

The blackish, pseudo-filamentous Silurian structures probably represent thermally altered, kerogenized remains of coccoid cyanobacterial mats. The Silurian deposits in the Bardo Mountains were influenced by Caledonian and/or Variscan thermal events⁸,

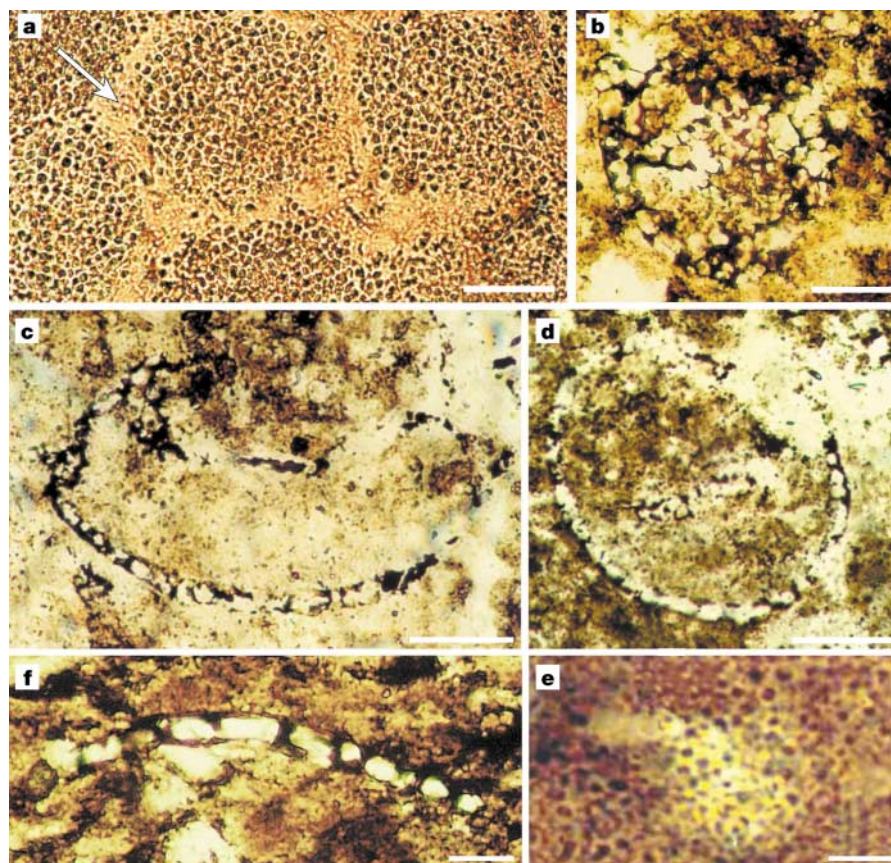


Figure 1 Light micrographs of mats of modern and early Silurian coccoid cyanobacteria. **a**, Modern mat of benthic coccoid cyanobacteria from Sulejów Dam, central Poland; the mat is composed of globular colonies surrounded by thick mucous envelopes (arrow). **b–f**, Examples of variously thermally altered colonies of Silurian coccoid cyanobacteria from Żdanów in the Bardo Mountains, Poland; these were originally composed of minute cells (magnified in **e**). Note the quasi-regular, segmented, blackish structures of thermal origin, which are strikingly similar to the purported Archaeal Apex Chert microfossils. Scale bars: **a**, 50 μm ; **b–d**, **f**, 20 μm ; **e**, 5 μm .

as evidenced by generations of microcracks filled with hydrothermally precipitated microcrystalline chalcedony and quartz. The cyanobacterial remains distributed in the vicinity of these cracks have been markedly altered, but still preserve traces of their primary biological structure.

Thermal alteration was apparently more advanced in the Apex Chert samples, leaving only isolated fragments in the chert background of much-changed ('carbonized') kerogen, and preserved pseudo-filamentous ghosts of the original biostructures. The generation of gaseous and bituminous hydrocarbons associated with thermal maturation and the conversion of kerogenous materials⁹ could have been responsible for the apparent reduction in volume and partial relocation of the cyanobacterial material.

Several inferences can be drawn from our observations of chertified and slightly thermally altered early-Silurian cyanobacterial mats. First, the early Archaeal Apex Chert filaments may have originated through late diagenetic (thermal or thermobaric) *in situ* alteration of kerogenized remnants of mucilage sheaths, capsules and extracellular polymer substances that originally enveloped groups of coccoid cells.

Second, the Apex Chert microfossil-like filamentous structures could therefore be biogenic but may represent diagenetic ghosts of benthic mats composed of colonial microorganisms resembling some modern chroococcalean cyanobacteria. Third, what have been described from the Apex Chert as 11 filamentous microbial taxa¹ may rather represent remnants of a homogeneous (and most probably monospecific) microbial community, similar to modern benthic coccoid cyanobacteria, that is also known from later Precambrian and Phanerozoic strata.

Józef Kaźmierczak, Barbara Kremer

Institute of Paleobiology, Polish Academy of Sciences, 00818 Warszawa, Poland
e-mail: jkaz@twarda.pan.pl

- Schopf, J. W. *Science* **260**, 640–646 (1993).
- Dalton, R. *Nature* **417**, 782–784 (2002).
- Brasier, M. D. et al. *Nature* **416**, 76–81 (2002).
- Wyżga, B. *Geologia Sudetica* **22**, 119–145 (1987).
- Kremer, B. in *Early Palaeozoic Palaeogeographies and Biogeographies of Western Europe and North Africa* (eds Alvaro, J. J. & Servais, T.) 36 (Univ. Sci. Technol., Lille, 2001).
- Komárek, J. & Anagnostidis, K. *Arch. Hydrobiol./Algolog. Stud.* **43**, 157–226 (1986).
- Horodyski, R. J. & Vonder Haar, S. J. *Sedim. Petrol.* **45**, 894–906 (1975).
- Aleksandrowski, P., Kryza, R., Mazur S., Pin, C. & Zalasiewicz, J. A. *Trans. R. Soc. Edinb.* **90**, 127–146 (2000).
- Tissot, B. P., Pelet, R. & Ungerer, Ph. *Am. Assoc. Petrol. Geol. Bull.* **71**, 1445–1466 (1987).

Supplementary information accompanies this communication on Nature's website.

Drag reduction through self-similar bending of a flexible body

Silas Alben*, Michael Shelley* & Jun Zhang*†

* Applied Mathematics Laboratory, Courant Institute of Mathematical Sciences, New York University, New York City, New York 10012, USA

† Department of Physics, New York University, New York City, New York 10003, USA

The classical theory of high-speed flow¹ predicts that a moving rigid object experiences a drag proportional to the square of its speed. However, this reasoning does not apply if the object in the flow is flexible, because its shape then becomes a function of its speed—for example, the rolling up of broad tree leaves in a stiff wind². The reconfiguration of bodies by fluid forces is common in nature, and can result in a substantial drag reduction that is beneficial for many organisms^{3,4}. Experimental studies of such flow–structure interactions⁵ generally lack a theoretical interpretation that unifies the body and flow mechanics. Here we use a flexible fibre immersed in a flowing soap film to measure the drag reduction that arises from bending of the fibre by the flow. Using a model that couples hydrodynamics to bending, we predict a reduced drag growth compared to the classical theory. The fibre undergoes a bending transition, producing shapes that are self-similar; for such configurations, the drag scales with the length of self-similarity, rather than the fibre profile width. These predictions are supported by our experimental data.

Experiments that cleanly reveal the nature of interactions between deformable bodies and flows are difficult to perform. Complications include controlling three-dimensional flow effects and visualizing the flows. Soap film is a convenient experimental system described by two-dimensional hydrodynamics in many aspects^{6–10}. Our soap-film flow tunnel is illustrated in Fig. 1. Driven by gravity, soapy water (1.5% Dawn dish detergent; density $\rho = 1 \text{ gm cm}^{-3}$) leaves an elevated reservoir and spreads into a vertical soap film (thickness $f = 1\text{--}3 \mu\text{m}$) descending between two straight nylon lines (tunnel width, 9.0 cm). Adjusting reservoir efflux rate adjusts flow velocity U through the range 0.5–3.0 m s^{-1} ; breakage occurs at velocities outside this range.

Half-way down the tunnel, a thin, flexible glass fibre (length $L = 1\text{--}5 \text{ cm}$; diameter, $34 \mu\text{m}$; rigidity $E = 2.8 \text{ erg cm}$), glued at its midpoint to a thin rod, is inserted transverse to the flow. We measure the drag force on the fibre, record its shape, and visualize the flow structures using interferometry, all as a function of flow speed. For comparison, a much more rigid fibre is also used ($L = 2.0 \text{ cm}$; $E = 2,000 \text{ erg cm}$). On the basis of fibre lengths, flow velocities, and soap film viscosity ν , the Reynolds number $\text{Re} = LU/\nu$ is typically large, in the range 2,000–40,000.

Figure 2a and b shows the flow patterns around a flexible fibre at two different flow speeds. Typical of high-Reynolds-number flows past bluff bodies, they are characterized by a thin separated boundary layer which divides the wake—containing slow-moving (a few cm s^{-1}), turbulent flow in which vorticity and viscosity are important—from the rest of the flow field, which is fast and laminar. But unlike rigid-body flows, as the flow speed is increased the body changes its shape by bending and so presents a smaller profile to the flow. Figure 3a shows in alternating colour the successively more-streamlined fibre shapes as flow speed is increased. We expect this to lead to a reduced growth of drag with flow speed, and this is borne out by Fig. 4a, which compares the drag induced by a nearly rigid fibre, well-approximated by U^2 growth, with that for a flexible fibre, which shows a much decreased, more slowly growing drag. An upper bound for the contribution of skin friction to the drag can be estimated by the expression for viscous drag per unit width on a flat

plate aligned with the flow¹, $1.33\rho U^2 L \text{Re}^{-1/2}$. For the parameters of our experiment, this value is at least an order of magnitude below the total drag; hence form drag predominates.

As we shall argue, a natural non-dimensional control parameter that scales linearly with flow velocity is:

$$\eta = \left(\frac{\rho f L^2 U^2 / 2}{E/L} \right)^{1/2} = (L/L_0)^{3/2} \text{ with } L_0 = (2E/\rho f U^2)^{1/3} \quad (1)$$

Here, η involves a ratio of fluid kinetic energy to elastic potential energy, or the ratio of fibre length to an intrinsic ‘bending length’ L_0 , which captures the competition between fibre rigidity and fluid forcing. Loaded elastic bodies often involve intrinsic length scales, such as the ‘buckling length’ of a beam under compression¹¹. To study the scaling of drag with flow speed, we introduce the drag coefficient $C_D = \text{Drag}/(\rho U^2 L f / 2)$, which for a rigid object in inviscid flow depends only on its geometric shape¹. A non-dimensional drag is defined as $D = C_D \eta^2$ so that D scales with η as the drag scales with velocity. Figure 4b plots D versus η . The data for a nearly rigid fibre (green) show approximately η^2 growth. The other data sets are for flexible fibres of equal rigidity but different lengths. The data for the longest fibre (blue) deviate at lower flow velocities, probably induced by proximity to the channel walls, but as the fibre profile narrows with increasing η , the data sets overlap. It appears that there is a transition at $\eta \approx O(1)$ from η^2 scaling of drag to a new, more slowly growing form.

To interpret the experiment, we abstract its crucial features and build a mathematical model coupling hydrodynamics to body flexibility. The fibre is modelled as a thin, inextensible elastic beam loaded by the difference in fluid pressure p between its upstream and downstream sides. The pressure is found by constructing exact, steady solutions to the inviscid, incompressible Euler equations, using free-streamline theory¹² (FST). This method, based upon conformal mapping, originated¹³ as a way to model flows around flat plates with a downstream wake. It has since been extended as a numerical method to compute flows around curved shapes, and to determine the shapes of sails under constant

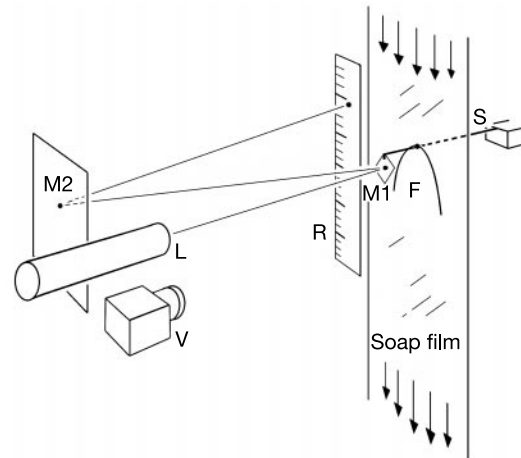


Figure 1 The layout of the experiment. A glass fibre (F) is inserted into a flowing soap-film tunnel (partly shown). The fibre is supported by a thin stainless-steel rod (S), which is clamped at one end. Fluid drag force acting on the fibre deflects this support slightly downwards. After a calibration using a known force, the drag force is determined by measuring the displacement of a laser beam (L) reflected from a small mirror (M1) mounted on S and then from a fixed mirror (M2). The rigidity of the support and the distances between the mirrors and the ruler (R) determine the overall sensitivity of the measurement. A video camera (V) records the position of the laser beam, which gives the force measurement. The soap solution is seeded with TiO_2 particles, allowing the flow speed to be measured by a laser Doppler velocimeter (not shown), aimed at the midline 7 cm above the filament. (Objects and distances are not drawn to scale).

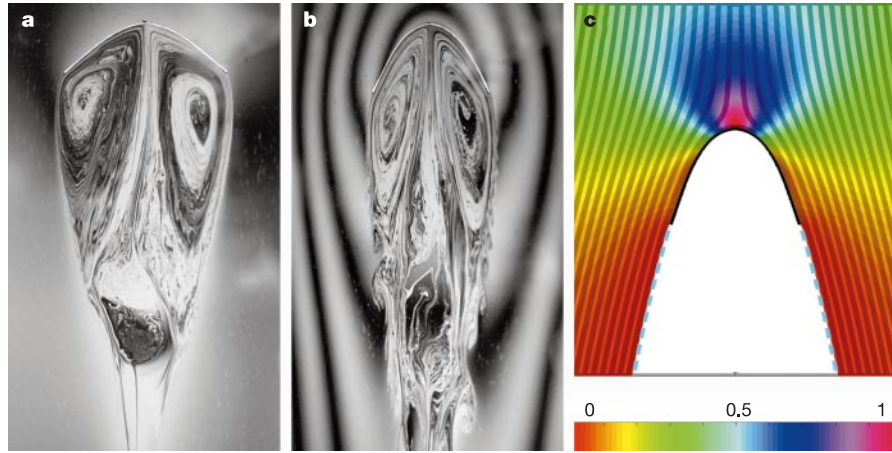


Figure 2 Examples of the flow field and fibre shape in experiment and model. **a,b**, Flow around a flexible fibre of length 4.1 cm and rigidity 2.8 erg cm, at flow speeds of 69 cm s^{-1} (**a**) and 144 cm s^{-1} (**b**). Monochromatic light from a low-pressure sodium lamp reflects from the two surfaces of the film, creating an interference pattern which

shows small variations in film thickness. **c**, Free-streamline model solution for non-dimensional velocity $\eta = 27$. The pressure field, shown in colour, varies from its maximum value of 1 at the stagnation point on the fibre (black) to its minimum value of 0 on the free streamlines (dashed lines); overlaid are the streamlines of the flow.

tension¹⁴. The extension here is to simultaneously determine the body and flow as a balance of elastic forces with pressure forces. By assuming incompressibility, we neglect possible effects of flow compressibility due to thickness variations in the soap film. Preliminary studies of such effects indicate that compressibility is not a dominant influence on the flow^{15,16}. In our case, this assumption seems to give a good account of the data.

The wake is represented by a stagnant region of constant-pressure fluid behind the body. Outside the wake, the flow is steady, inviscid and irrotational. 'Free streamlines' separate from the fibre ends, dividing these two regions. The main flow is a potential flow, with the boundary conditions $\mathbf{u} \cdot \mathbf{n} = 0$ on the fibre, and $|\mathbf{u}| = U$ on the free streamlines, where $\mathbf{u}$ is fluid velocity, $\mathbf{n}$ the unit normal to the fibre, and U the flow speed at infinity. We take the simplest version of FST, which has an infinite wake and fluid pressure equal to wake pressure at the free streamlines. We set the far-field pressure p_∞ to zero. The formulation of ref. 17 is convenient, relating $\mathbf{u}$ to the body shape given as $\theta(s)$, the tangent angle as a function of the signed arc length s along the fibre.

Given $\mathbf{u}$, the pressure loading is given by the Bernoulli equation: $[p] = p_{\text{fibre}} - p_{\text{wake}} = p_{\text{fibre}} - p_\infty = \rho(U^2 - |\mathbf{u}_{\text{fibre}}|^2)/2$. Here p_{wake} is the far-field pressure, or zero, and hence the pressure jump across the fibre is p_{fibre} . In real flows, p_{wake} is nearly constant but significantly less than zero near the body, rising gradually to zero downstream. For this reason, the infinite-wake model must be modified to accurately represent the pressure distribution and drag on the body. This has usually been achieved with a more complicated wake structure which takes p_{wake} as an input parameter, generally determined by empirical considerations¹⁸. We retain the infinite-wake model with $p_{\text{wake}} = 0$, but instead apply a uniform multiplicative increase to the free-stream velocity in the model over that in the corresponding experiment. Through Bernoulli's equation, $[p]$ is then increased in a way similar to that achieved with a non-zero p_{wake} . This technique has produced good empirical results for wake models¹⁹.

The pressure load is balanced by the fibre's tensile and bending forces:

$$-(Ts)' + (E\kappa'\mathbf{n})' = fp_{\text{fibre}}\mathbf{n} \quad (2)$$

with T the axial tension, $\kappa = \theta'$ the fibre curvature, and $\mathbf{s}$ its unit tangent vector. The superscript $'$ denotes differentiation with respect to arc length. Equation (2) is supplemented by 'free-end' boundary conditions: $T = \kappa = \kappa' = 0$ at fibre ends, $s = \pm L/2$. The

fibre is fixed and clamped at its midpoint (that is, $\theta(0) = 0$).

Making the beam/FST system non-dimensional simplifies equation (2) to a nonlinear scalar equation for beam curvature:

$$\kappa'' + \frac{1}{2}\kappa^3 = \eta^2 p_{\text{fibre}} \quad (3)$$

(A similar expression involving the curvature arises, for example, in studying the deformations of an inextensible elastic beam in a very viscous fluid²⁰.) This closes the model, with the non-dimensional speed η its only parameter. The beam/FST system is solved numerically (S.A., M.S. and J.Z., manuscript in preparation).

A solution is found for each value of η , giving fluid velocity and pressure fields, fibre shape, and free streamline shapes. Figure 2c

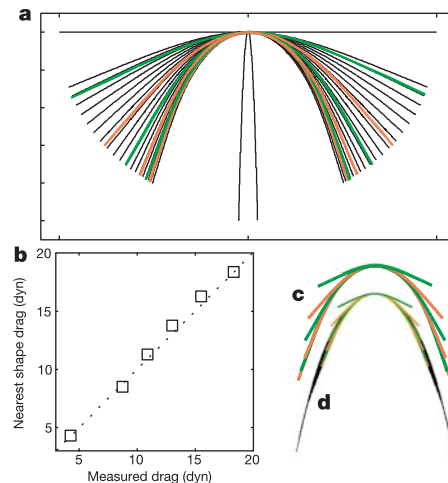


Figure 3 Comparison of six experimental fibre shapes with model shapes. **a**, Shapes from the experiment (green and orange lines) superimposed on a manifold of numerical solutions (black lines). The numerical solutions range from $\eta = 5.5$ to 33, in increments of 10%. The flat plate ($\eta = 0$) and a very bent solution ($\eta = 30,000$) are shown for comparison. **b**, Comparison of measured drag on experimental fibres in **a** with computed drag of nearest numerical solutions, determined by matching tip curvature. **c,d**, Transition to self-similarity in experimental (**c**) and numerical solutions (**d**) (not shown at the same scale). The six fibres, and their nearest numerical solutions, are dilated by $\eta^{2/3}$ and superimposed. The black fibres are numerical solutions at higher η , and show the subsequent convergence to a universal shape, as predicted by the model.

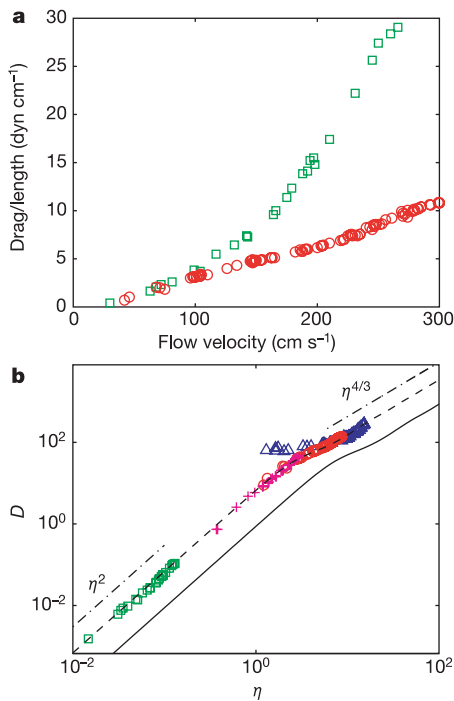


Figure 4 Comparison of drag data from experiment and model. **a**, Drag per unit fibre length versus flow velocity for a flexible fibre ($L = 3.3$ cm; red circles) and a rigid fibre ($L = 2.0$ cm; green squares). **b**, Log-log plot of drag data in **a**, non-dimensionalized as $D = C_D \eta^2$; also shown are data for two fibres with the same rigidity as the flexible fibre in **a** but different lengths (pink plus signs, $L = 1.8$ cm; blue triangles, $L = 5.2$ cm). The solid line is the values of D in the model (well-fitted by power laws η^2 for $\eta \ll 1$ and $\eta^{4/3}$ for $\eta \gg 1$, as shown by dash-dotted lines). The short-dashed line is the solid line shifted by a factor of 2.8 in η , to correspond with the shift in the shape comparison. This shift compensates for the ‘back pressure’ in the wake, as explained in the text.

shows a detailed solution, and Fig. 3a shows computed fibre shapes for η small to large. For $\eta \ll 1$, the fibre is nearly straight, but as η becomes $O(1)$ there is a sharp transition to bending. The most notable property of the transition is an emerging self-similarity in fibre shape. For $\eta \gg 1$, large curvature becomes confined to an ever-smaller region near the tip as η increases. The length of this ‘tip region’ is proportional to the bending length, $L_0 = L\eta^{-2/3}$. This sets the scale for self-similarity: when dilated by $\eta^{2/3}$ the shapes overlap a universal, quasi-parabolic form (Fig. 3d). The transition is also captured in the computed dimensionless drag D (Fig. 4b). For $\eta \ll 1$ the model’s drag scales quadratically, as does the experimental data; for $\eta \approx O(1)$, the model’s drag undergoes a transition to a much slower growth, scaling as $\eta^{4/3}$ for $\eta \gg 1$.

Self-similarity of shape gives an interpretation for the $\eta^{4/3}$ drag-scaling. Rescaling arc length as $S = s\eta^{2/3}$, curvature as $\eta^{2/3}K(s\eta^{2/3}) = \kappa(s)$, and pressure as $P(s\eta^{2/3}) = p(s)$, equation (3) becomes $K^3/2 + K'' = P$, with η appearing only in the boundary conditions: $K = K' = 0$ at $S = \pm\eta^{2/3}/2$. As $\eta \rightarrow \infty$, K and P tend to the curvature and pressure of the universal shape. The drag is given by

$$D = \eta^2 \int_{\text{fibre}} [p] dy = \eta^{4/3} \int_{-\eta^{2/3}/2}^{+\eta^{2/3}/2} \left(\frac{1}{2} K^3 + K'' \right) \frac{dY}{dS} dS \quad (4)$$

where $Y = y\eta^{2/3}$. As $\eta \rightarrow \infty$, the integral converges to its value for the universal shape, and hence the drag scales as the prefactor $\eta^{4/3}$.

Figure 3a compares computed to experimental shapes for the full range of flow speeds. Aligning the midpoints, the relative displacement from the midpoint between the experimental shape and the nearest computed shape is less than five per cent over the entire

length. Figure 3c indicates the emergence of self-similarity in the experimental shapes, as in the computed shapes, though the upper limit on the soap-film flow speed impedes a thorough comparison for large η . In Fig. 3b the measured drag for the fibres is compared to that of their nearest computed solutions, demonstrating that the similar shapes in theory and experiment also give similar drags. We interpret this to mean that the model captures the nature of the pressure jump distribution across the fibre.

Through this comparison, we obtain a quantitative estimate for the multiplicative increase of η needed for the model to account for the wake pressure and its effect on the drag. The values of η for the corresponding theoretical and experimental fibre shapes of Fig. 3a differ by an almost uniform factor of 2.8. In Fig. 4b, the model and experiment show very similar drag scaling and transition when the theoretical curve is shifted leftward in η to adjust for this factor.

Fluid pressure and elastic bending forces seem sufficient to describe the drag reduction observed in our experiment. Though negligible for our flow speeds, skin friction may become important in much faster flows, where the fibre would be more swept-back, like a flat plate aligned with the flow. As implied earlier, the viscous drag on the fibre would then scale as $U^{3/2}$, eventually dominating the $U^{4/3}$ form drag.

A final comment concerns the relation of bending length to drag. Two-dimensional form drag usually scales as $\rho U^2 W$, where W is profile width. In our case, the parabolic form of the universal shape implies that W decreases as $\eta^{-1/3}$. Hence one would expect $D \sim \eta^{5/3}$. Instead we find that the drag is induced on the bending length $L_0 \sim \eta^{-2/3}$, giving the more slowly growing form $D \sim \eta^{4/3}$. In essence, the tip region creates a parabolic wake in which the rest of the body sits, contributing little to the drag. $\square$

Received 13 August; accepted 14 October 2002; doi:10.1038/nature01232.

1. Batchelor, G. K. *An Introduction to Fluid Dynamics* (Cambridge Univ. Press, Cambridge, 1967).
2. Vogel, S. Drag and reconfiguration of broad leaves in high winds. *J. Exp. Bot.* **40**, 941–948 (1989).
3. Denny, M. Extreme drag forces and the survival of wind-swept and water-swept organisms. *J. Exp. Biol.* **194**, 97–115 (1994).
4. Koehl, M. How do benthic organisms withstand moving water? *Am. Zool.* **24**, 57–70 (1984).
5. Vogel, S. *Life in Moving Fluids* (Princeton Univ. Press, Princeton, 1994).
6. Couder, Y., Chomaz, J. M. & Rabaud, M. On the hydrodynamics of soap films. *Physica D* **37**, 384–405 (1989).
7. Gharib, M. & Derango, P. A liquid-film (soap film) tunnel to study two-dimensional laminar and turbulent shear flows. *Physica D* **37**, 406–416 (1989).
8. Goldburg, W. I., Rutgers, M. A. & Wu, X.-L. Experiments on turbulence in soap films. *Physica A* **239**, 340–349 (1997).
9. Rutgers, M. A., Wu, X.-L. & Daniel, W. B. Conducting fluid dynamics experiments with vertically falling soap films. *Rev. Sci. Instrum.* **72**, 3025–3037 (2001).
10. Zhang, J., Childress, S., Libchaber, A. & Shelley, M. Flexible filaments in a flowing soap film as a model for one-dimensional flags in a two-dimensional wind. *Nature* **408**, 835–839 (2000).
11. Segel, L. *Mathematics Applied to Continuum Mechanics* (MacMillan, New York, 1977).
12. Birkhoff, G. & Zarantonello, E. H. *Jets, Wakes, and Cavities* (Academic, New York, 1957).
13. Helmholtz, H. Über discontinuierliche Flüssigkeitsbewegungen. *Monatsber. Berlin Akad.* 215–268 (1868); reprinted in *Phil. Mag.* **36**, 337–346 (1868).
14. Dugan, J. P. A free-streamline model of the two-dimensional sail. *J. Fluid Mech.* **42**, 433–446 (1970).
15. Rivera, M., Vorobieff, P. & Ecke, R. E. Turbulence in flowing soap films: velocity, vorticity, and thickness fields. *Phys. Rev. Lett.* **81**, 1417–1420 (1998).
16. Rivera, M. & Wu, X.-L. External dissipation in driven two-dimensional turbulence. *Phys. Rev. Lett.* **85**, 976–979 (2000).
17. Hureau, J., Brunon, E. & Legallais, P. Ideal free streamline flow over a curved obstacle. *J. Comput. Appl. Math.* **72**, 193–214 (1996).
18. Wu, T. Y. Cavity and wake flows. *Annu. Rev. Fluid Mech.* **4**, 243–284 (1972).
19. Smith, F. T. Laminar flow of an incompressible fluid past a bluff body: the separation, reattachment, eddy properties and drag. *J. Fluid Mech.* **92**, 171–205 (1979).
20. Goldstein, R. E. & Langer, S. A. Nonlinear dynamics of stiff polymers. *Phys. Rev. Lett.* **75**, 1094–1097 (1995).

Acknowledgements We thank F. Vollmer, S. Childress, A. Libchaber and P. Palffy-Muhoray for conversations, and Y.-Q. Xia who assisted in a preliminary experimental study of the fibre/flow system. M.S. thanks A. Gorieli, who originally suggested this as an interesting problem. This work was supported by the National Science Foundation and the Department of Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.S. (e-mail: shelley@cims.nyu.edu).

Multiple ionization of atom clusters by intense soft X-rays from a free-electron laser

H. Wabnitz*, L. Bittner*, A. R. B. de Castro†, R. Döhrmann*, P. Gürtler*, T. Laarmann*, W. Laasch*, J. Schulz*, A. Swiderski*, K. von Haefen*‡§, T. Möller*, B. Faatz*, A. Fateev‡§, J. Feldhaus*, C. Gerth*, U. Hahn*, E. Saldin‡, E. Schneidmiller‡, K. Sytchev‡, K. Tiedtke*, R. Treusch* & M. Yurkov||

* Hamburger Synchrotronstrahlungslabor HASYLAB at Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22603 Hamburg, Germany

† LNLS, 13084-971, R. Guiseppe Maximo Scolfaro, Campinas SP, Brazil, and IFGW-UNICAMP, 13083-970 Campinas SP, Brazil

‡ Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22603 Hamburg, Germany

|| Joint Institute for Nuclear Research, Dubna, 141980 Moscow Region, Russia

Intense radiation from lasers has opened up many new areas of research in physics and chemistry, and has revolutionized optical technology. So far, most work in the field of nonlinear processes has been restricted to infrared, visible and ultraviolet light¹, although progress in the development of X-ray lasers has been made recently². With the advent of a free-electron laser in the soft-X-ray regime below 100 nm wavelength³, a new light source is now available for experiments with intense, short-wavelength radiation that could be used to obtain deeper insights into the structure of matter. Other free-electron sources with even shorter wavelengths are planned for the future. Here we present initial results from a study of the interaction of soft X-ray radiation, generated by a free-electron laser, with Xe atoms and clusters. We find that, whereas Xe atoms become only singly ionized by the absorption of single photons, absorption in clusters is strongly enhanced. On average, each atom in large clusters absorbs up to 400 eV, corresponding to 30 photons. We suggest that the clusters are heated up and electrons are emitted after acquiring sufficient energy. The clusters finally disintegrate completely by Coulomb explosion.

Short-wavelength radiation, namely vacuum-ultraviolet (VUV) radiation and soft and hard X-rays, is a very powerful tool for studying the properties of matter—such as geometrical and electronic structure, chemical composition and magnetic structure. The ability to ionize matter allows direct insight into electronic structure and element composition because the binding energies of electrons are characteristic of the individual elements. Diffraction of hard X-rays is very important because it allows for structure determination in complex systems with atomic resolution. Free-electron lasers (FELs) for soft X-rays based on the principle of self-amplification of spontaneous emission (SASE) are now becoming available for experiments. These lasers combine the advantages of short-wavelength sources (like synchrotron radiation) with those of lasers; for example, short pulse length, high peak power, coherence and diffraction-limited collimation⁴. Short-wavelength X-ray FELs are therefore expected to open the way to new types of experiments^{5,6}, such as the study of chemical reactions in real time with element selectivity⁷, or structure determination of single biomolecules⁸.

Here we report the study of the interaction of intense soft X-rays from an FEL with matter. Xe atoms and clusters are chosen because they can be ionized even by single 12.7-eV photons (the ionization potential of Xe atoms is 12.1 eV). The experiments were performed by irradiating atoms and clusters with ~100-fs-long FEL pulses at 98 nm wavelength and a power density of up to $7 \times 10^{13} \text{ W cm}^{-2}$.

The resulting ions are detected with a time-of-flight (TOF) mass spectrometer.

TOF mass spectra for different cluster sizes recorded at $2 \times 10^{13} \text{ W cm}^{-2}$ are shown in Fig. 1 (experimental details are given in the legend). The most striking result is the surprisingly different ion signal from atom and cluster beams. Whereas only singly charged ions are observed after irradiation of isolated atoms, atomic ions with charges up to 8+ are detected if clusters are irradiated. Doubly charged ions resulting from ionization of isolated Xe atoms are not detected above the noise level of ~1%. Clusters absorb many photons, and completely disintegrate into singly and multiply charged ions. The mass peaks are very broad, indicating that the ions have a high kinetic energy. This can be understood in terms of a Coulomb explosion process. The population of different ion states and their kinetic energy strongly depends on the power density. This is shown in Fig. 2 for clusters comprising 1,500 atoms. At the highest power level of $7 \times 10^{13} \text{ W cm}^{-2}$, charge states up to 8+ are detected. Lowering the power density strongly reduces the intensity of highly charged ions. At $2 \times 10^{11} \text{ W cm}^{-2}$, only singly charged Xe ions with high kinetic energy can be detected. The high kinetic energy of the ions

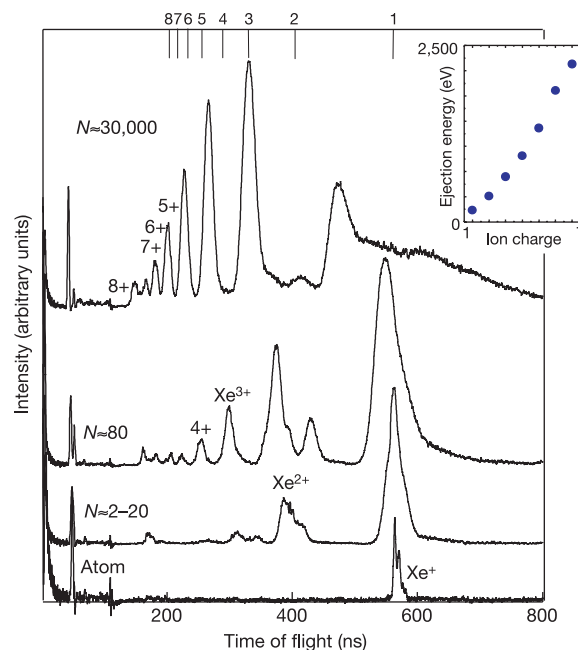


Figure 1 Time-of-flight (TOF) mass spectra of ionization products of Xe atoms and clusters. The spectra were recorded after ionization with soft X-rays (98 nm wavelength) at an average power density of $2 \times 10^{13} \text{ W cm}^{-2}$. The atomic spectrum (bottom trace) shows a splitting into several lines owing to the different isotopes. After irradiation of clusters, highly charged ions are observed. The mass peaks are rather broad and displaced with respect to the calculated flight times indicated by thin vertical lines (different charge states) in the uppermost part of the figure. This indicates that the ions have high kinetic energies. *N* is the number of atoms per cluster. Inset, the kinetic energy of ions as a function of the charge for *N* = 1,500. The experiments are performed as follows. A pulsed beam of atoms or clusters was prepared by expanding Xe gas at high pressure (0.1–30 bar) through a small conical nozzle (0.1 mm diameter, 15° half-angle). The composition of the beam and the size of the clusters could be controlled by varying the gas pressure^{28,29}. The cluster beam passed through a small skimmer into the main chamber. Soft X-rays from the FEL⁴ with 98 nm wavelength (bandwidth 0.5 nm) were focused with an elliptical mirror onto the atom or cluster beam. The pulse energy was measured with a carefully calibrated channel-plate detector, which detects a well-defined percentage of FEL light scattered from a wire. The diameter of the focal spot was ~20 μm. The power density is calculated assuming a pulse length of 100 fs. The ions produced in the interaction zone were detected with a TOF detector mounted perpendicular to the polarization direction and with a digital sampling oscilloscope operating at a 1-Hz repetition rate.

§ Present address: Lehrstuhl für Physikalische Chemie II, Ruhr-Universität Bochum, D-44780 Bochum, Germany.

and the absence of molecular fragments (Xe_2^+ , Xe_3^+ , Xe_4^+ , ...) indicates that the clusters completely disintegrate even at reduced power density. Further reducing the intensity by lowering the gain of the FEL results in a decreasing signal and simultaneously a change of the kinetic energy of Xe^+ (lowest curve in Fig. 2).

The strong dependence on the power density is a clear sign that optical nonlinear processes dominate the ionization of the clusters at the power levels used. If the power density is increased from 2×10^{11} to $7 \times 10^{13} \text{ W cm}^{-2}$, the mean charge state increases from 1 to ~ 2.5 . Therefore, we conclude that at the highest power density each atom in the cluster of 1,500 atoms loses on average 2–3 electrons. The average kinetic energy per ion is derived from the measured flight time in Figs 1 and 2. It varies strongly with the charge state and the cluster radius. For Xe^+ the kinetic energy increases almost linearly with the cluster radius, whereas for high charge states it starts to saturate for large clusters (H.W., manuscript in preparation). For Xe^{7+} , kinetic energies of more than 2 keV were observed. In small clusters, the kinetic energy increases quadratically with the charge state (see Fig. 1 inset). This behaviour is a clear signature of a Coulomb explosion⁹. From the results presented in Fig. 1 we conclude that an energy of up to several hundred eV per atom is taken from the FEL beam.

Coulomb explosion of clusters is not a new phenomenon¹⁰. Clusters irradiated with infrared light at $>10^{16} \text{ W cm}^{-2}$ completely disintegrate^{11–13}. The interaction of intense lasers with clusters has attracted great interest, and has been extensively studied experimentally^{14,15} and theoretically^{16–21} in the past few years. The interpretation of the heating mechanism and of the explosion dynamics, especially if driven by the thermal expansion of the electron gas, is still a subject of discussion⁹. On the other hand, it is agreed that the experimental findings are explained by field

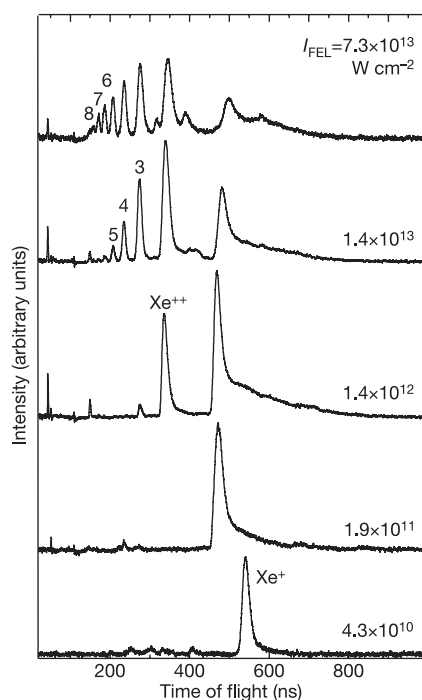


Figure 2 TOF mass spectra recorded after irradiation of 1,500-atom Xe clusters as a function of power density. The power density (I_{FEL}) is given next to each trace. The spectrum at the bottom was recorded at a reduced gain of the FEL. The intensity of highly charged ions increases with increasing power density. Experimental details: the intrinsic pulse energy of the ~ 100 -fs-long SASE-FEL pulses⁴ typically varied between 1.5 and 25 μJ . The spectra with power densities of (1.9×10^{11}) – $(7 \times 10^{13}) \text{ W cm}^{-2}$ were taken with pulses of 25 μJ energy. The power density could be lowered to $10^{10} \text{ W cm}^{-2}$ at 1.5 μJ by moving the cluster beam out of the focus.

ionization of the atoms and the clusters by the strong electric field of the infrared laser^{18,20}.

The effects observed here when clusters are exposed to short-wavelength radiation are somewhat surprising, because Coulomb explosion starts at $10^{11} \text{ W cm}^{-2}$; this is much lower than the power density needed to induce Coulomb explosion in the infrared. Moreover, the frequency ω of the FEL is so high that the direct effect of the laser field on the electron movement is expected to be small²². Whether field ionization is efficient can be estimated from textbook formulas²³. Field ionization by the laser field dominates if the Keldysh parameter γ is smaller than 1. Under our conditions, γ varies between 8 (Xe atoms, ionization potential $I_p = 12.1 \text{ eV}$) and ~ 100 (electrons bound to highly charged Xe clusters) and therefore field ionization can hardly account for the observations. Consequently, additional processes have to be considered for the understanding of absorption, ionization and explosion dynamics with short-wavelength radiation. Three questions have to be answered. (1) What is the ionization mechanism that causes the massive electron ejection? (2) How can the high charge states of the

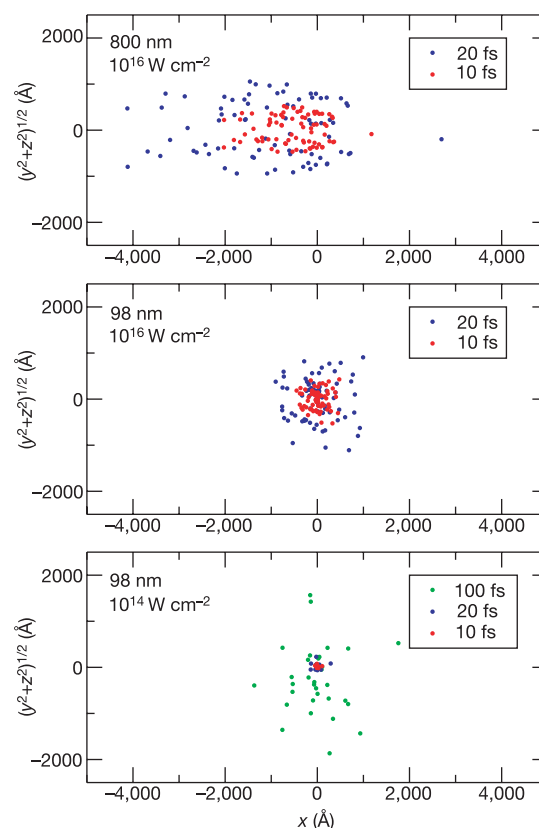


Figure 3 Classical simulations of the electron motion and ionization for Xe_{13} clusters after irradiation with infrared radiation and soft X-rays. Electron trajectories were calculated for 100-fs-long pulses of constant intensity. The position of the electrons is shown 10, 20 and 100 fs after the onset of a light pulse. The position of the ions was that of the neutral cluster and kept fixed during the simulation. The radiation field is polarized in the x direction. The vertical scale $(y^2 + z^2)^{1/2}$ shows the vertical distance of the electrons from the cluster. The uppermost plot corresponds to typical conditions for infrared lasers (800 nm wavelength, power density $10^{16} \text{ W cm}^{-2}$). Most electrons leave the cluster during the first optical cycle into the polarization direction of the light owing to field ionization. At 98 nm wavelength, the electrons are emitted isotropically by thermionic emission. Details of the calculation will be published elsewhere (H.W., manuscript in preparation). Further calculations were performed assuming that at least one electron per atom is unbound owing to inner ionization by the FEL photons. The number of unbound electrons is varied between 1 and 6 per atom in view of the high charge states seen in the experiment. At 98 nm wavelength and a power density of $1 \times 10^{14} \text{ W cm}^{-2}$, up to 29 electrons can leave the cluster.

fragments and their high kinetic energy be explained? (3) Which process allows the cluster to absorb such large amounts of energy?

To help the understanding of these findings, we recall that the binding energy of electrons in clusters increases strongly with the charge of the cluster. The total energy needed to remove N electrons from a cluster comprising N atoms is approximately proportional to N^2 , and can be as large as 10^7 eV for $N = 30,000$ if the radius of the cluster remains unchanged during ionization. We assume that the ionization of the cluster is a two-step process, as in the case of irradiation with infrared light^{17,20}. In a first step, atoms inside the cluster lose at least one electron, which then can move almost freely inside the cluster. This process is called inner ionization. In a second step, the electrons are removed from the cluster to infinity. This process is called outer ionization. Inner ionization of the electrons can be easily explained, because the energy of the FEL photons exceeds the ionization energy of Xe atoms. Therefore, we assume that valence electrons in the cluster are promoted to excited states (conduction band). At a power level of a few times 10^{13} W cm⁻², this takes place in the first few femtoseconds of the FEL pulse because the cross-section is large²⁴ (30–50 Mbarn).

To explain the outer ionization, we have performed simulations of the electron movement in Xe₁₃ and Xe₅₅ clusters using a classical model similar to that in ref. 17 in order to get insight into the absorption and ionization processes. The motion of unbound electrons inside the clusters is simulated under the action of the forces of the ions and the electrons in the cluster and of the FEL field. The results are presented in Fig. 3. At 98 nm wavelength and a power density of 10^{14} W cm⁻² or 10^{16} W cm⁻², the electrons are emitted isotropically, whereas at the typical conditions of optical laser experiments (800 nm, 10^{16} W cm⁻²) the electrons are preferentially emitted into the polarization direction of the electric field of the laser. The latter is a clear signature of field ionization. The isotropic emission of electrons at 98 nm shows that field ionization

does not play a role. A close examination of the electron trajectories shows (H.W., manuscript in preparation) that the electrons are scattered many times before they leave the cluster. The clusters are heated, and electrons escape once sufficient energy is acquired. The steps of the proposed ionization process are illustrated in Fig. 4. We assume that highly charged ions are formed at the cluster surface by field ionization in the strong Coulomb field of the ions inside the clusters^{16,18}. The electric field at the surface of large highly charged Xe clusters calculated by Coulomb's law is typically $10\text{--}50$ V Å⁻¹. This is up to 20 times larger than the maximum field strength of the FEL at 10^{14} W cm⁻². 30 V Å⁻¹ are sufficient²⁵ to produce Xe⁸⁺, giving a straightforward explanation for the presence of high charge states.

We now discuss further the mechanisms that allow the cluster to absorb such a high amount of energy. The energy released in the Coulomb explosion corresponds to the absorption of many photons. This number is estimated by dividing the average kinetic energy of the ejected ions by the photon energy. For clusters comprising 1,500 atoms, approximately 400 eV per atom are absorbed at 7×10^{13} W cm⁻², corresponding to ~ 30 photons per atom. The absorption of 30 photons at 7×10^{13} W cm⁻² corresponds to an average cross-section of 10 Mbarn if we assume that the photons are absorbed sequentially. The absorption can not be explained by simple ionization of Xe atoms. Although 100 photons per shot fall into the absorption cross-section (50 Mbarn) for the initial ionization, the energy of a single photon is not sufficient to excite or ionize further Xe ions inside the cluster. Therefore, other absorption mechanisms have to be considered.

The standard inverse bremsstrahlung model describes successfully the absorption of infrared light by ionized clusters. Our estimates (H.W., manuscript in preparation) for 98 nm wavelength based on the work in ref. 12 give only 15 eV per atom for singly charged ions, which is substantially lower than the experimentally derived value. At 98 nm the absorption by inverse bremsstrahlung is inefficient because the ponderomotive energy is small (<150 meV), and the oscillation amplitude of electrons heating the cluster is only 0.01 nm and thus much smaller than the diameter of the clusters (1–5 nm). Our classical simulations for Xe₅₅ in the geometry of the neutral cluster predict an absorption of 30 eV per atom in 100 fs if an average charge state 1 per atom is considered. It increases to 85 eV per atom for an average charge state of 2 per atom. In agreement with recent theoretical work¹⁷, the simulations show a strong wavelength-dependent resonance behaviour that is due to oscillations of unbound electrons in the entire cluster and to plasmon absorption. In contrast to experiments with infrared lasers, under our conditions the plasma frequency is always smaller than the laser frequency. For Xe clusters the plasma frequency shifts from 2.7 eV for singly charged atoms to 6.6 eV for six charges per atom. Even if a strong broadening of the plasmon absorption due to damping of the electron motion by collisions is considered, the absorption is approximately 5–10 times less efficient than experimentally observed (H.W., manuscript in preparation). Other approaches, such as Brunel absorption²⁶, also cannot account for the strong absorption because the FEL frequency is much larger than the plasma frequency.

Two different effects may account for the efficient absorption. The model calculation gives absorption rates that depend on cluster size. According to recent theoretical work²⁷, the absorption of a cluster plasma confined by a strong Coulomb potential is enhanced compared with that of an infinite plasma. This enhancement compared with conventional inverse bremsstrahlung is due to collisions of the electrons with the whole cluster ion²⁷. Whether this effect can explain the enhanced absorption could be checked by calculations for larger clusters. Furthermore, we note that a classical modelling of the absorption process is limited, because enhancements due to resonant intermediate states are not included. The photon energy of the FEL is of the same order of magnitude as the

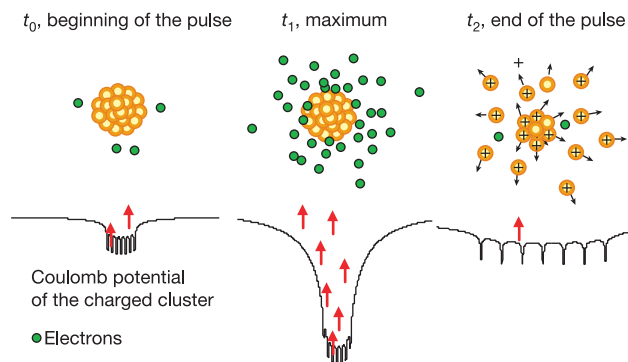


Figure 4 Schematic illustration of the ionization of Xe clusters and a subsequent Coulomb explosion. The Coulomb potential of the clusters is shown at the beginning of the FEL pulse, at its maximum and at the end of the pulse. Photons are indicated by arrows. Ionization proceeds in steps. In the first step—depending on the size of the cluster—a few electrons are directly ejected from the cluster because their binding energy is smaller than the energy of the FEL photons. However, once the charge of the cluster has reached a critical value, the binding energy of the electrons becomes larger than the photon energy. At that stage, only hot electrons in excited states can leave the cluster, but electrons in the valence band and in excited states can absorb photons. We assume that after a few femtoseconds a nanoscale plasma is formed, which is confined by the Coulomb potential of the positive charge in the cluster. The radius of the electron cloud can be considerably larger than the cluster radius because the Coulomb potential is long range. As a result we expect that the electrons have a high probability of being outside the cluster. These loosely bound electrons can be removed from the cluster by the absorption of additional photons or by collisions with other electrons in the presence of ions close by. Once a high number of electrons have left, the clusters start to disintegrate as a result of the Coulomb repulsion of the ions ('Coulomb explosion'). At the same time outer ionization becomes easier because the Coulomb potential gets weaker¹⁸.

level spacing in Xe atoms and Xe ions. It is therefore quite likely that a quantum mechanical description of the absorption process which includes resonant intermediate states could give larger absorption than the classical model.

In conclusion, absorption of short-wavelength radiation and subsequent ionization in clusters differs considerably from that in the optical spectral range. Absorption and ionization start by single-photon absorption as described by quantum mechanics. After many unbound electrons are created, a plasma is formed. Our experimental results give evidence that absorption in such cluster plasma is stronger than predicted by calculations with classical models. We assume that quantum-mechanical modelling is needed at these short wavelengths to explain the efficient energy deposition seen in the experiment. On the other hand, the classical simulation clearly shows that electrons can leave the cluster by a photon-assisted thermionic emission. Field ionization, the dominant ionization process at optical frequencies, does not contribute to cluster ionization. □

Received 11 July; accepted 7 October 2002; doi:10.1038/nature01197.

1. Batani, D., Joachain, C. J., Martellucci, S. & Chester, A. N. *Atoms, Solids, and Plasmas in Super-Intense Laser Fields* (Kluwer Academic, New York, 2001).
2. Smith, R. & Key, M. H. *X-Ray Lasers 2000* (eds Jamelot, L. G., Möller, C. & Klinsnick, A.) 383–388 (EDP Science, Saint Malo, 2000).
3. Andruszkow, J. *et al.* First observation of self-amplified spontaneous emission in a free-electron laser at 109 nm wavelength. *Phys. Rev. Lett.* **85**, 3825–3828 (2000).
4. Ayvazyan, V. *et al.* Generation of GW radiation pulses from a VUV free-electron laser operating in the femtosecond regime. *Phys. Rev. Lett.* **88**, 104802 (2002).
5. Åberg, T. *et al.* *DESY Report TESLA FEL 95-03* (DESY, Hamburg, 1995).
6. Materlik, G. & Tschentscher, T. (eds) *DESY Report TESLA FEL 2001-05* (DESY, Hamburg, 2001).
7. Shenoy, G. K. & Stöhr, J. *LCLS: The First Experiments* (SLAC, Stanford, 2000).
8. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Haidu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).
9. Russek, M., Lagarde, H. & Blenski, T. Cluster explosion in an intense laser pulse: Thomas-Fermi model. *Phys. Rev. A* **63**, 13203 (2000).
10. Synder, E. M., Buzza, S. A. & Castleman, A. W. Jr Intense field matter interactions: Multiple ionisation of clusters. *Phys. Rev. Lett.* **77**, 3347–3350 (1996).
11. Ditmire, T. *et al.* High-energy ions produced in explosions of superheated atomic clusters. *Nature* **386**, 54–56 (1997).
12. Ditmire, T., Donnelly, T., Rubenchik, A. M., Falcone, R. W. & Perry, M. D. Interaction of intense pulses with atomic clusters. *Phys. Rev. A* **53**, 3379–3402 (1996).
13. Lezius, M., Dobosz, S., Normand, D. & Schmidt, M. Explosion dynamics of rare gas clusters in strong laser fields. *Phys. Rev. Lett.* **80**, 261–264 (1998).
14. Köller, L. *et al.* Plasmon-enhanced multi-ionization of small metal clusters in strong femtosecond laser fields. *Phys. Rev. Lett.* **82**, 3783–3786 (1999).
15. Shao, Y. L. *et al.* Multi-keV electron generation in the interaction of intense laser pulses with Xe clusters. *Phys. Rev. Lett.* **77**, 3343–3346 (1996).
16. Rose-Petruck, C., Schafer, K. J., Wilson, K. R. & Barty, C. P. J. Ultrafast electron dynamics and inner-shell ionization in laser driven clusters. *Phys. Rev. A* **55**, 1182–1190 (1996).
17. Last, I. & Jortner, J. Quasiresonance ionization of large multicharged clusters in a strong laser field. *Phys. Rev. A* **60**, 2215–2221 (1999).
18. Last, I. & Jortner, J. Dynamics of the Coulomb explosion of large clusters in a strong laser field. *Phys. Rev. A* **62**, 13201 (2000).
19. Brewczyk, M., Clark, C. W., Lewenstein, M. & Rzaewski, K. Stepwise explosion of atomic clusters induced by a strong laser field. *Phys. Rev. Lett.* **80**, 1857–1860 (1998).
20. Krainov, V. P. & Roshchupin, A. S. Dynamics of Coulomb explosion of large Xe clusters irradiated by a super-intense ultra-short laser pulse. *J. Phys. B* **34**, L297–L303 (2001).
21. McPherson, A., Thompson, B. D., Borisov, A. B., Boyer, K. & Rhodes, C. K. Multiphoton-induced X-ray emission at 4–5 keV from Xe atoms with multiple core vacancies. *Nature* **370**, 631–634 (1994).
22. Brewczyk, M. & Rzaewski, K. Over-the-barrier ionization of multi electron atoms by intense VUV free-electron laser. *J. Phys. B* **32**, L1–L4 (1999).
23. Delone, N. B. & Krainov, V. P. *Multiphoton Process in Atoms* (ed. Lambropoulos, P.) 1–3 (Springer, Berlin, 2000).
24. Samson, J. A. R. New energy levels in xenon and krypton. *Phys. Lett.* **8**, 107–108 (1964).
25. Aust, S., Strickland, D., Meyerhofer, D. D., Chin, S. L. & Eberly, J. H. Tunneling ionization of noble gases in a high-intensity laser field. *Phys. Rev. Lett.* **63**, 2212–2215 (1989).
26. Brunel, F. Not-so resonant, resonant absorption. *Phys. Rev. Lett.* **59**, 52–55 (1987).
27. Kostyukov, I. Y. Inverse-bremsstrahlung absorption of an intense laser field in cluster plasma. *JETP Lett.* **73**, 393–397 (2001).
28. Hagena, O. F. Condensation in free jets: Comparison of rare gases and metals. *Z. Phys. D* **4**, 291–299 (1987).
29. Karnbach, R., Joppien, M., Stapelfeldt, J., Wörmer, J. & Möller, T. CLULU: An experimental setup for luminescence measurements on van der Waals clusters with synchrotron radiation. *Rev. Sci. Instrum.* **64**, 2838–2849 (1993).

Acknowledgements We thank the TTF team at DESY, especially P. Castro, M. Minty, D. Nölle, H. Schlarb and S. Schreiber, for running the accelerator; we also thank J. R. Schneider for support and discussions. The first group of authors (H.W. to T.M.) built the apparatus for the cluster experiment and performed the experiment; the second group of authors (B.F. to M.Y.) worked on the FEL and the diagnostics. We thank K.H. Meiwes-Broer, T. Brabec, C. Rose-Petruck, J. Krzywinski, M. Lezius, I. Kostyukov, J.M. Rost, E. Rühl, U. Saalman, J. Jortner and M. Smirnov and their research groups for discussions and comments, and J. Sutter for critically reading the manuscript. This work was supported by the DFG.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.M. (e-mail: thomas.moeller@desy.de).

Non-framework cation migration and irreversible pressure-induced hydration in a zeolite

Yongjae Lee*, Thomas Vogt*, Joseph A. Hriljac†, John B. Parise‡, Jonathan C. Hanson§ & Sun Jin Kim||

* *Physics Department, Brookhaven National Laboratory, Upton, New York 11973, USA*

† *School of Chemical Sciences, University of Birmingham, Birmingham B15 2TT, UK*

‡ *Geosciences Department, State University of New York, Stony Brook, New York 11794, USA*

§ *Chemistry Department, Brookhaven National Laboratory, Upton, New York 11973, USA*

|| *Nano-Materials Research Center, Korea Institute of Science and Technology, Seoul 130-650, Korea*

Zeolites crystallize in a variety of three-dimensional structures in which oxygen atoms are shared between tetrahedra containing silicon and/or aluminium, thus yielding negatively charged tetrahedral frameworks that enclose cavities and pores of molecular dimensions occupied by charge-balancing metal cations and water molecules¹. Cation migration in the pores and changes in water content associated with concomitant relaxation of the framework have been observed in numerous variable-temperature studies^{2–5}, whereas the effects of hydrostatic pressure on the structure and properties of zeolites are less well explored^{6–8}. The zeolite sodium aluminosilicate natrolite was recently shown to undergo a volume expansion at pressures above 1.2 GPa as a result of reversible pressure-induced hydration⁹; in contrast, a synthetic analogue, potassium gallosilicate natrolite, exhibited irreversible pressure-induced hydration with retention of the high-pressure phase at ambient conditions¹⁰. Here we report the structure of the high-pressure recovered phase and contrast it with the high-pressure phase of the sodium aluminosilicate natrolite. Our findings show that the irreversible hydration behaviour is associated with a pronounced rearrangement of the non-framework metal ions, thus emphasizing that they can clearly have an important role in mediating the overall properties of zeolites.

The pressure-induced changes of the unit cell parameters of a powder sample of the potassium gallosilicate form of natrolite (K-GaSi-NAT) reveal a volume expansion of about 1.0% between 1.2(1) and 1.7(1) GPa (numbers in parentheses are errors on measured pressures) (Fig. 1a)¹⁰. Structurally, this expansion differs from that observed in the sodium aluminosilicate form of natrolite (Na-ALSi-NAT): in K-GaSi-NAT, the *c*-axis, along which the rigid T₅O₁₀ (T = Al, Si, Ga, and so on) tetrahedral building units join to form a chain (Fig. 1b), expands by about 0.4%, whereas it

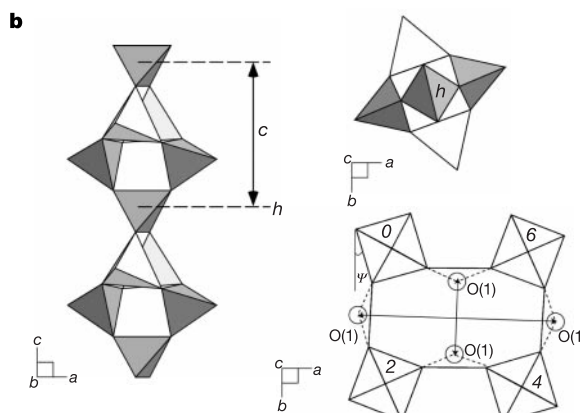
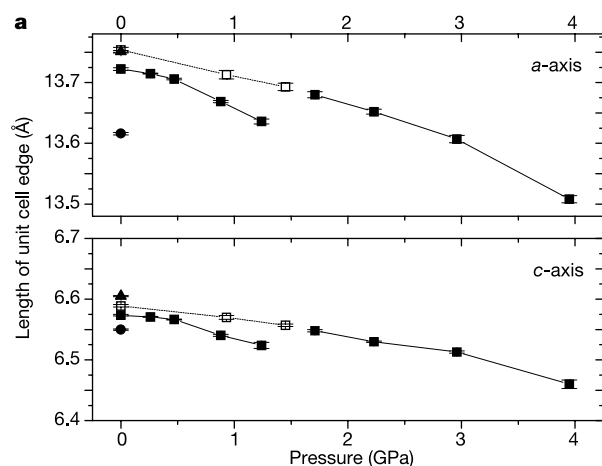


Figure 1 Structural features of the natrolite framework and its pressure response. **a**, Evolution of the unit-cell edge lengths (Å) with pressure for K-GaSi-NAT. Squares represent increase (filled) and release (unfilled) of pressure from the *in situ* measurements. Triangles and circles indicate measurements at room temperature and -160°C , respectively, on a single crystal recovered from 1.9 GPa. **b**, A polyhedral representation of the chain found in the NAT framework shown along [010] (left) and [001] (upper right). The repeat distance of the T_5O_{10} building unit of five Ga/Si- (or Al/Si-) tetrahedra constitutes the c -axis length (c). Lower right, a simplified skeletal

representation of the channel along the c -axis. The height (h) of the central tetrahedral node (upper right) can be described in terms of eighths of the repeat distance, and natrolite shows a 2460 -type connectivity of the neighbouring chains. Vertices represent Ga/Si- (or Al/Si-) tetrahedral atoms, and oxygen atoms are omitted. The overall rotation angle of the chains, ψ , is defined by the (average) angle between the sides of the quadrilateral around the chain and the a - (and b -) axis. The O(1) oxygen atoms are shown to permit visualization of the channel openings in the GaSi-NAT framework.

contracts continuously in Na-AlSi-NAT (ref. 9). The pressure-induced expansion in K-GaSi-NAT is therefore three-dimensional, but is essentially two-dimensional in Na-AlSi-NAT. More strikingly, the larger-volume phase of K-GaSi-NAT endures after pressure release at ambient conditions (Fig. 1a), with the unit cell volume of the recovered sample being 0.8% larger (more than 40σ) than that

measured before compression. Diffraction data were collected 3 and 8 days after pressure release and, within 5σ , no indication of a further volume contraction was detected.

Because the powder form of the high-pressure phase of K-GaSi-NAT can be retained at ambient conditions, a successful attempt was made to produce single crystals (see Methods). The structure was

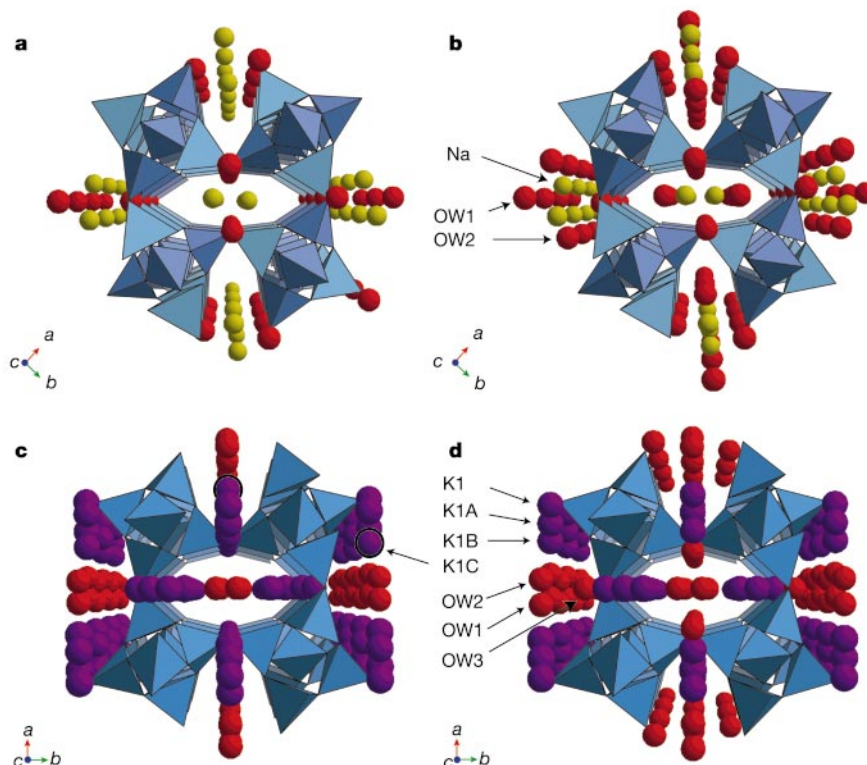


Figure 2 Polyhedral representations of Na-AlSi-NAT and K-GaSi-NAT before and after pressure-induced hydration. **a**, Na-AlSi-NAT at 0.40 GPa; **b**, Na-AlSi-NAT at 1.51 GPa; **c**, K-GaSi-NAT as synthesized; **d**, K-GaSi-NAT recovered from 1.9 GPa. Tetrahedra in

Na-AlSi-NAT are shown in two colours to illustrate the ordering of Al/Si over the framework tetrahedral sites, whereas Ga/Si in K-GaSi-NAT are disordered and shown in one colour. K1C atoms are emphasized with bold outlines (see the text).

determined *ex situ*, and the crystal structures of K-GaSi-NAT before and after pressure-induced hydration (PIH) are shown in Fig. 2 and compared with the structures of Na-AlSi-NAT from an *in situ* pressure experiment^{9,10}. The water contents of both K-GaSi-NAT and Na-AlSi-NAT double after PIH; for K-GaSi-NAT the refined unit-cell compositions before and after the pressure experiment are $K_{7.5(3)}Ga_{8.0(1)}Si_{12.0(1)}O_{40} \cdot 6.3(6)H_2O$ and $K_{7.9(5)}Ga_8Si_{12}O_{40} \cdot 12.24(16)H_2O$, respectively (Supplementary Table 1). After PIH, a new partly occupied water site OW3 appears in close proximity to the statistically distributed potassium cation sites. This causes the potassium cations, distributed over K1, K1A, K1B and K1C sites, to migrate away from the centre of the channel and merge into three sites. The occupancies of the initial water sites OW1 and OW2 increase by 64% and 35%, respectively. All water molecules, including those located at the new OW3 site, coordinate to the potassium cations, with interatomic distances ranging from 2.45(12) to 3.35(13) Å (Supplementary Table 2). As a consequence, the average

potassium-to-framework oxygen distance range increases from 2.661(4)–2.794(9) Å to 2.71(1)–2.92(2) Å. The pressure-induced expansion is thus related to an overbonding of the potassium cations.

Another structural response during PIH is the decrease in the overall rotation angle of the chains, ψ , from 17.5(1)° to 17.0(1)° (see Fig. 1b). This change in ψ with pressure is similar to, although smaller than, the corresponding change observed in Na-AlSi-NAT and suggests that PIH is coupled to the relaxation of the overall framework distortion by expanding the pore space across the channel; in fact, the opening of the channel, defined by the shortest and longest interatomic distances between two chain-bridging oxygens (O(1); see Fig. 1b), increases from 5.70(1) Å × 9.54(1) Å to 5.80(1) Å × 9.57(1) Å before and after PIH, respectively. However, the degree of overall chain rotation is much less than those observed in the aluminosilicate natrolite (23.7(1)° to 26.4(1)°). The reason for this is attributed to a combined effect of the different non-framework cation distributions and the increased flexibility of

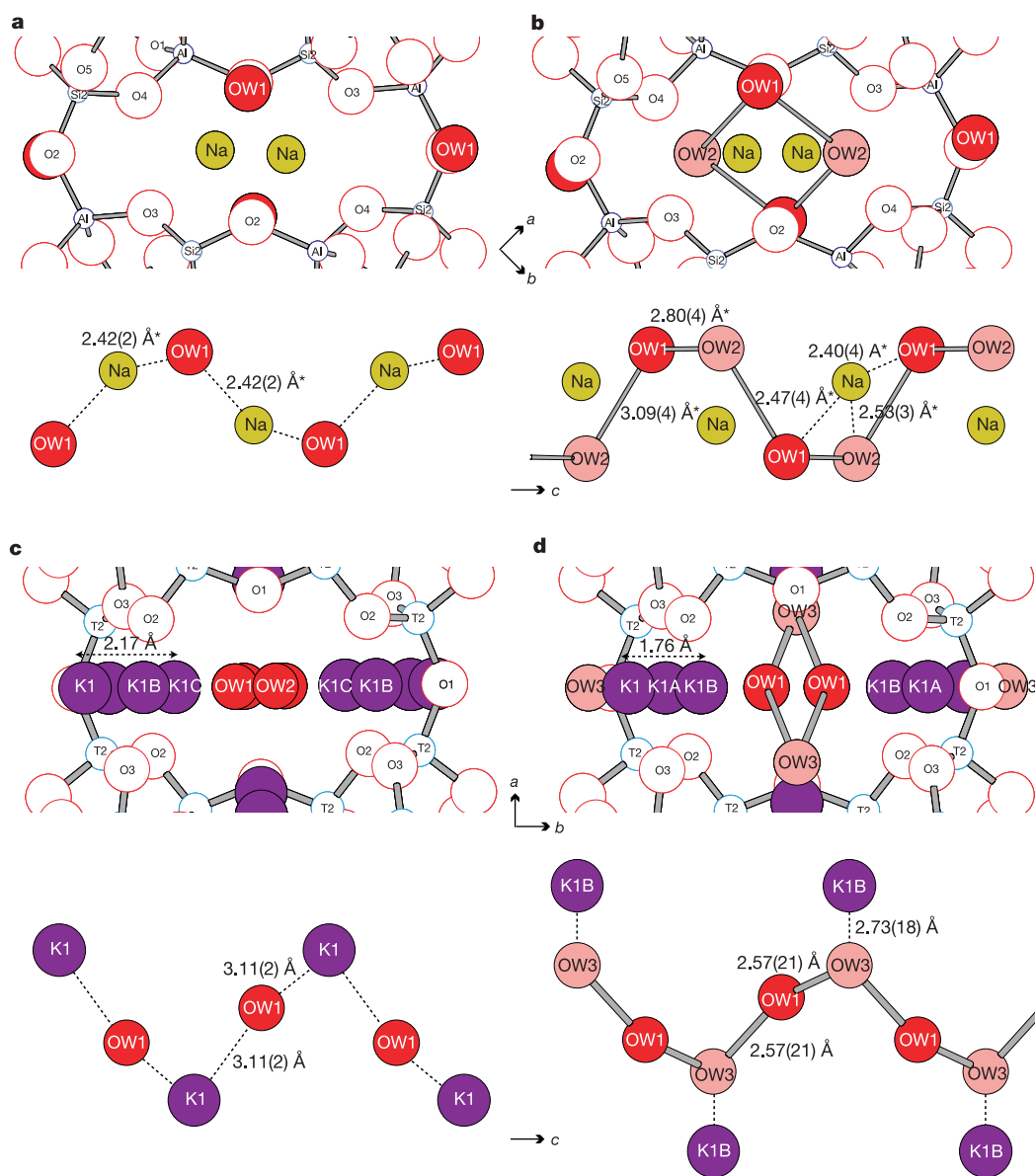


Figure 3 Expanded views of the channel contents of Na-AlSi-NAT and K-GaSi-NAT before and after pressure-induced hydration. **a**, Na-AlSi-NAT at 0.40 GPa; **b**, Na-AlSi-NAT at 1.51 GPa; **c**, K-GaSi-NAT as synthesized; **d**, K-GaSi-NAT recovered from 1.9 GPa. Each lower figure shows a view perpendicular to the channel axis. Hydrogen bondings of water nanostructures (**b** and **d**) are emphasized with thick lines, and interactions between water

and non-framework cations are shown with dotted lines. New water sites are emphasized in light red. For K-GaSi-NAT, the water molecules at the statistical OW2 site are not shown for clarity (lower **c** and **d**) and one of the potassium sites is chosen to emphasize the cation migration (lower **c** and lower **d**). Structural details of Na-AlSi-NAT were taken from ref. 10.

the T–O–T angles in K–GaSi–NAT than in Na–AlSi–NAT. The latter might also explain the three-dimensional swelling unique to the gallosilicate natrolite in contrast to the two-dimensional swelling of Na–AlSi–NAT. Given the greater flexibility of the T–O–T angle in the gallosilicate framework, the incorporation of additional water molecules into the channels at high pressure would allow the T–O–T angles within the chain (T–O(2)–T and T–O(3)–T) to relax together with the angles between the chains (T–O(1)–T; see Fig. 1b and Supplementary Table 2).

In an attempt to decrease the static crystallographic disorder of the non-framework species in the PIH state of K–GaSi–NAT as well as to explore the reversibility back to the original state, variable-temperature experiments were performed with the same quenched single crystal. The cell parameters measured at room temperature (24 °C) after 1 month at ambient conditions did not show any changes (within 2σ) from the values measured immediately after the pressure release, confirming again the long-term retention of pressure-induced hydration at ambient conditions. The temperature was then decreased to –160 °C with a stream of dry N₂ blowing onto the crystal. However, it was found that under these conditions the material lost its excess water, vacating the OW3 site, and the potassium cations shifted back towards the vacant water site away from the channel walls (Supplementary Table 1). The occupancy of the K1 site was 40% less than in the original state before PIH, and the K1C site was not reoccupied when losing the excess water. This shows that the potassium redistribution was not completely reversible even when the PIH was reversed. As the temperature was increased to room temperature (24 °C) again, an unknown small-volume phase formed. This new phase persisted after heating to 100 °C. Work is in progress to investigate in greater detail the structural evolutions of both phases of K–GaSi–NAT as a function of temperature.

It has been previously pointed out that PIH in Na–AlSi–NAT results in the formation of new hydrogen-bonded water nanotubes, which enclose sodium cations (Figs 2b and 3b)^{9,10}. The PIH in K–GaSi–NAT, in contrast, leads to the formation of hydrogen-bonded water nanochains, which are retained at ambient conditions (Figs 2d and 3d); the hydrogen bonding between the water molecules at the OW3 site and those at the OW1 (or OW2) site generates a zigzag array of water molecules along the zeolitic channel (Fig. 3d). Unlike the water nanotubes in Na–AlSi–NAT, the water nanochains in K–GaSi–NAT have finite lengths because the water sites are not fully occupied (Supplementary Table 1). The different formation of water nanostructures after PIH can be related to the different arrangement of non-framework cations. At ambient conditions, potassium cations in K–GaSi–NAT occupy sites bound by the T₁₀O₂₀ windows close to the channel walls (Figs 2c and 3c) and water molecules are found along the channels (OW1 and OW2). This is the reverse of the sodium and water arrangements found in Na–AlSi–NAT (Figs 2a and 3a). The PIH in Na–AlSi–NAT does not cause any major redistribution of non-framework cations, whereas the potassium cations in K–GaSi–NAT are rearranged and merged into three closely separated sites away from the channel. The water nanochain in K–GaSi–NAT is thus surrounded by potassium cations (Figs 2d and 3d), whereas the water nanotube in Na–AlSi–NAT encloses sodium cations (Figs 2b and 3b). The migration and rearrangement of non-framework cations under pressure might have a crucial role in stabilizing the hydrogen-bonded water nanochains so that the PIH state in K–GaSi–NAT can be retained at ambient conditions but not that in Na–AlSi–NAT. As similar crystal chemical features occur in other zeolites and related framework materials, irreversible PIH and other unusual properties are likely to be observed in future work. Irreversible PIH, although at lower pressures, has the potential to immobilize tritiated water or other pollutant molecules. Furthermore, the expanded pore openings in the PIH state might markedly facilitate the ion exchange properties of these dense pore zeolites, which might then be used to

trap radioactive ions by a subsequent pressure release ('trap-door' mechanism). □

Methods

In situ high-pressure experiments

The synthesis and crystal structure of K–GaSi–NAT were reported previously¹¹. Variable-pressure powder diffraction data were measured with diamond anvil cells at beamline X7A of the National Synchrotron Light Source (NSLS). Each powdered sample was loaded into a 200-μm-diameter sample chamber in a pre-indented stainless steel gasket, along with a few small ruby chips as a pressure gauge. A methanol:ethanol:water mixture (16:3:1 by volume) was used as a pressure medium (hydrostatic up to ~10 GPa)¹². The pressure at the sample was measured by detecting the shift in the R1 emission line of the included ruby chips. No evidence of non-hydrostatic conditions or pressure anisotropy was detected during our experiments, and the instrumental errors on the pressure measurements ranged between 0.05 and 0.1 GPa. Typically, the sample was equilibrated for ~15 min at each measured pressure, and diffraction data were collected for 3–5 h (3–35° 2θ, λ = 0.6942(1) Å) using a micro-focused (~200 μm) monochromatic X-ray provided by a bent Si (220) monochromator and a gas-proportional position-sensitive detector¹³. The sample pressure was raised in 0.5–1.0 GPa increments before subsequent data measurements up to 5 GPa. There was no evidence of stress-induced peak broadening or pressure-induced amorphization. Several sets of diffraction data were measured after gradual pressure release. Unit-cell parameters were determined by whole-pattern fitting with the LeBail method. The diffraction peaks were modelled by varying only a half-width parameter in the pseudo-Voigt profile function. Bulk moduli were calculated by fitting the Murnaghan equation of state to normalized volumes ($V/V_0 = [1 + B'P/B_0]^{-1/B'}$, where $B' = (\partial B/\partial P)_{P=0} = 4$). The structural evolution of Na–AlSi–NAT was determined from the variable-pressure powder diffraction data and the Rietveld structure refinement^{9,10}.

Ex situ high-pressure experiments

Several needle-shaped crystals of K–GaSi–NAT were loaded into a diamond anvil cell as described above and initially pressurized at 1.9 GPa. After 2 weeks the sample pressure was measured as 1.3 GPa. The sample was subsequently recovered from the diamond anvil cell and kept at ambient conditions for 3 d before diffraction data measurement. A needle-shaped crystal of 5 × 5 × 50 μm³ was mounted on a glass fibre, and data collection was performed at ambient conditions at beamline X7B of the NSLS with an imaging plate detector (Mar345; 2,300 × 2,300 pixels). Diffraction data were measured in an oscillation mode by rotating the crystal in φ by 4° in 5 min per frame; 40 frames were measured. The intensities were integrated and merged with the DENZO program¹⁴. Corrections for Lorentz polarization effects were made, and no absorption correction was applied. The space group was determined by examining the systematic absences. Structural models were determined and refined with the SHELX program¹⁵. Refinements were based on full-matrix least-squares techniques on F^2 . Anisotropic displacement factors were included for all framework atoms, whereas two sets of restrained isotropic displacement parameters were used for disordered non-framework cations and water molecules. The agreement factors for the final model are $R_1 = 0.0621$, $wR_2 = 0.1462$, goodness of fit = 2.2 using 251 unique reflections with $I > 2\sigma$ ($R_{\text{int}} = 4.6\%$). The refinement results including atomic parameters and selected bond distances and angles are listed in Supplementary Information.

Received 27 May; accepted 29 October 2002; doi:10.1038/nature01265.

1. Breck, D. W. *Zeolite Molecular Sieves* (Krieger, Malabar, Florida, 1984).
2. Baur, W. H. & Joswig, W. The phases of natrolite occurring during dehydration and rehydration studied by single crystal x-ray diffraction methods between room temperature and 923 K. *Neues Jb. Miner. Mh.* **4**, 171–187 (1996).
3. Miyamoto, T., Katada, N., Kim, J. H. & Niwa, M. Acidic property of MFI-type gallosilicate determined by temperature-programmed desorption of ammonia. *J. Phys. Chem. B* **102**, 6738–6745 (1998).
4. Lee, Y. *et al.* New insight into cation relocations within the pores of zeolite rho: In situ synchrotron X-ray and neutron powder diffraction studies of Pb- and Cd-exchanged rho. *J. Phys. Chem. B* **105**, 7188–7199 (2001).
5. Kuznicki, S. M. *et al.* A titanosilicate molecular sieve with adjustable pores for size-selective adsorption of molecules. *Nature* **412**, 720–724 (2001).
6. Hazen, R. M. Zeolite molecular-sieve 4A—anomalous compressibility and volume discontinuities at high-pressure. *Science* **219**, 1065–1067 (1983).
7. Belitsky, I. A., Fursenko, B. A., Gubada, S. P., Kholdeev, O. V. & Seryotkin, Y. V. Structural transformations in natrolite and edingtonite. *Phys. Chem. Minerals* **18**, 497–505 (1992).
8. Lee, Y. *et al.* Phase transition of zeolite rho at high-pressure. *J. Am. Chem. Soc.* **123**, 8418–8419 (2001).
9. Lee, Y., Hriliac, J. A., Vogt, T., Parise, J. B. & Artioli, G. First structural investigation of a super-hydrated zeolite. *J. Am. Chem. Soc.* **123**, 12732–12733 (2001).
10. Lee, Y., Vogt, T., Hriliac, J. A., Parise, J. B. & Artioli, G. Pressure-induced volume expansion of zeolites in the natrolite family. *J. Am. Chem. Soc.* **124**, 5466–5475 (2002).
11. Lee, Y., Kim, S. J. & Parise, J. B. Synthesis and crystal structures of gallium- and germanium-variants of the fibrous zeolites with the NAT, EDI and THO structure types. *Microporous Mesoporous Mater.* **34**, 255–271 (2000).
12. Hazen, R. M. & Finger, L. W. *Comparative Crystal Chemistry* (Wiley, New York, 1982).
13. Smith, G. C. X-ray imaging with gas proportional detectors. *Synch. Radiat. News* **4**, 24–30 (1991).
14. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
15. Sheldrick, G. M. Phase annealing in SHELX90—direct methods for larger structures. *Acta Crystallogr. A* **46**, 467–473 (1990).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank J. Hu and the Geophysical Laboratory of the Carnegie Institute for access to their ruby laser system at beamline X17C. This work was supported by a Laboratory Directed Research and Development grant from Brookhaven National Laboratory (BNL) (Pressure in Nanopores). J.H. acknowledges financial support from the Royal Society, and J.P. thanks NSF and the American Chemical Society—Petroleum Research Fund. Research performed in part at the NLSL at BNL is supported by the US DOE, Division of Materials Sciences.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.V. (e-mail: tvogt@bnl.gov).

A short-term sink for atmospheric CO₂ in subtropical mode water of the North Atlantic Ocean

Nicholas R. Bates*, A. Christine Pequignet*, Rodney J. Johnson* & Nicolas Gruber†

* Bermuda Biological Station For Research, Inc., Ferry Reach, Bermuda, GE01
† University of California at Los Angeles, Los Angeles, California 90095-4996, USA

Large-scale features of ocean circulation, such as deep water formation in the northern North Atlantic Ocean¹, are known to regulate the long-term physical uptake of CO₂ from the atmosphere by moving CO₂-laden surface waters into the deep ocean. But the importance of CO₂ uptake into water masses that ventilate shallower ocean depths, such as subtropical mode waters² of the subtropical gyres, are poorly quantified. Here we report that, between 1988 and 2001, dissolved CO₂ concentrations in subtropical mode waters of the North Atlantic have increased at a rate twice that expected from these waters keeping in equilibrium with increasing atmospheric CO₂. This accounts for an extra ~0.4–2.8 Pg C (1 Pg = 10¹⁵ g) over this period (that is, about 0.03–0.24 Pg C yr⁻¹), equivalent to ~3–10% of the current net annual ocean uptake of CO₂ (ref. 3). We suggest that the lack of strong winter mixing events, to greater than 300 m in depth, in recent decades is responsible for this accumulation, which would otherwise disturb the mode water layer and liberate accumulated CO₂ back to the atmosphere. However, future climate variability (which influences subtropical mode water formation^{1,4–8}) and changes in the North Atlantic Oscillation⁹ (leading to a return of deep winter mixing events) may reduce CO₂ accumulation in subtropical mode waters. We therefore conclude that, although CO₂ uptake by subtropical mode waters in the North Atlantic—and possibly elsewhere—does not always represent a long-term CO₂ sink, the phenomenon is likely to contribute substantially to interannual variability in oceanic CO₂ uptake³.

The net global ocean exchange of CO₂ between the ocean and the atmosphere depends on the rates of CO₂ flux across the air–sea interface, and the supply of CO₂ to (and its removal from) the ocean mixed layer. The transfer of CO₂ from the ocean mixed layer to deeper depths is a complex function of biological and physical processes that operate over a wide range of timescales (<1 to >1,000 yr). In subpolar/polar regions, the formation and subsequent sinking of intermediate and deep waters represent a physical mechanism for transporting CO₂ from the surface to the deep ocean. In the subtropical gyres, the shallow depths between the seasonal and main thermoclines are ventilated as a result of the formation of subtropical mode waters (STMWs)². However, the

rates and interannual variability in the uptake of CO₂ from the atmosphere into STMW and the fate of this CO₂ are poorly quantified at present.

The STMW of the North Atlantic Ocean is formed each winter by cooling and convective mixing at the northern edges of the subtropical gyre south of the Gulf Stream^{3–5}. The shallow depths of this gyre (~250–400 m)^{4,10} are ventilated during STMW formation. The STMW layer is found throughout the subtropical gyre, and its age increases following the mean geostrophic circulation from the site of STMW formation to the western boundary current¹¹. This water mass is classically defined by temperatures ranging from 17.8 to 18.4 °C (ref. 4), by a salinity of ~36.5 ± 0.05, and by a minimum in the vertical gradient of potential density (or isopycnal potential vorticity)^{2,5,7,12}. The strength and geographic extent of STMW formation is highly variable interannually^{4,5,10,13}, and primarily coupled to North Atlantic Oscillation (NAO)^{1,4,5,8,13} variability—the NAO being a dipole meridional oscillation in atmospheric pressure between the Iceland Low and the Azores High⁹. Observations of STMW variability date back to 1954 with time-series hydrographic data collected at the Hydrostation S site (32° 10' N, 64° 30' W) near Bermuda^{6,12–14}, expendable bathythermograph (XBT) surveys since the 1960s^{4,7}, and monthly sampling since 1988 at the US Joint Global Ocean Flux Study (JGOFS) Bermuda Atlantic Time-series Study (BATS) site (31° 50' N, 64° 10' W) near Bermuda¹⁵.

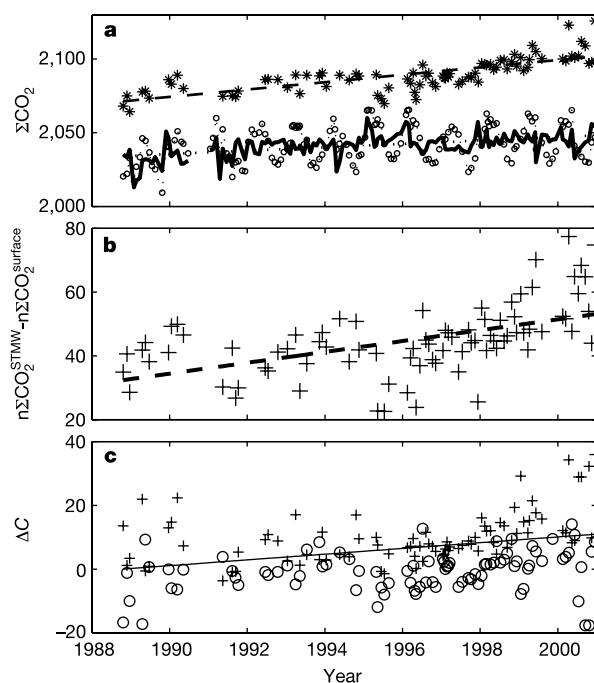


Figure 1 Interannual variability and trends of ΣCO_2 and hydrography in the western North Atlantic near Bermuda (1988–2001). **a**, Trends of surface ΣCO_2 (open circles) and STMW ΣCO_2 (asterisks) sampled at BATS from October 1988 to September 2001. Surface-layer ΣCO_2 data were normalized to the mean salinity of 36.6. Also shown are seasonally adjusted surface layer ΣCO_2 data, where the cyclo-stationary seasonal cycle has been removed by fitting harmonic functions of time with periods of 12, 6 and 4 months to the data. Regression lines are shown for the STMW layer (dashed line) and the surface layer (solid line). **b**, Difference between surface-layer and STMW ΣCO_2 sampled at BATS from October 1988 to September 2001. Both surface and STMW data have been adjusted to account for the cyclo-stationary seasonal cycle. The dashed line is the linear regression of the data. **c**, Long-term trends of C_{air} (regression line), the gas exchange component C_{aex} ('plus' symbols), and the biological component C_{bio} (open circles) in the STMW layer. Details and regression statistics are given in Table 1. $n\Sigma\text{CO}_2$, ΣCO_2 and ΔC units are $\mu\text{mol kg}^{-1}$.

Ocean inorganic carbon data have been collected at the BATS site each month since 1988. Water-column seawater samples from 158 cruises were analysed for total carbon dioxide (ΣCO_2) content by gas extraction and coulometric methods^{15,16}. Highly precise and accurate measurements of ΣCO_2 , a requisite for determining long-term trends, have been achieved by analysing seawater certified reference materials over time^{15,16}. ΣCO_2 samples for near-surface waters only have been collected at Hydrostation S since 1983¹⁷ and, unfortunately, do not provide comparative information about STMW CO_2 variability before 1988.

Long-term observations at the BATS site indicate that the ΣCO_2 concentrations of surface waters and of the STMW layer have increased at divergent rates over time (dC/dt) since sampling began in 1988. In the surface layer, the mean rate of increase of $n\Sigma\text{CO}_2$ (ΣCO_2 normalized to a constant salinity to account for the dilution/concentration effect of freshwater addition/removal) is $1.25 \pm 0.14 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (see Methods) (Fig. 1a; Table 1). The rate of change of surface ocean ΣCO_2 in the western North Atlantic is thus similar to that observed in the subtropical gyre of the North Pacific Ocean ($1.18 \pm 0.32 \mu\text{mol kg}^{-1} \text{CO}_2 \text{yr}^{-1}$)¹⁸. In STMW, however, the mean rate of change of ΣCO_2 is much higher, increasing at $2.64 \pm 0.26 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (using a 17.8 – 18.4°C temperature criterion to define STMW; Fig. 1a, Table 1). This result is independent of the criterion used to define the STMW layer, as we find a very similar rate of change of $2.22 \pm 0.27 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (Table 1) if we use a density criterion (that is, $26.4 \sigma_\theta$ isopycnal surface).

These long-term trends indicate that the subtropical gyre of the North Atlantic is at present acting as a significant and detectable sink for anthropogenic CO_2 . From 1988 to 2001, ΣCO_2 has increased in the surface layer and STMW by ~ 15 ($\sim 0.8\%$) and $\sim 33 \mu\text{mol kg}^{-1}$ ($\sim 1.5\%$), respectively, while the vertical difference (that is, difference between surface and STMW TCO_2 , or $\delta_z C$) has increased from ~ 35 to $\sim 55 \mu\text{mol kg}^{-1}$ (Fig. 1b). Before 1988, the mean vertical difference in ΣCO_2 between surface and STMW, determined from ΣCO_2 data from a few stations in the Sargasso Sea during the 1972–86 period, was $\sim 34 \mu\text{mol kg}^{-1}$ (Fig. 2a). The

vertical difference in ΣCO_2 should remain fairly constant over time if the physical and biological processes that influence the ΣCO_2 contents of the subtropical gyre were in steady state and the ocean was just accumulating anthropogenic CO_2 in response to the increasing atmospheric CO_2 concentrations. However, the large changes in $\delta_z C$ and in the ΣCO_2 contents of the STMW layer since the late 1980s indicate that the subtropical gyre of the North Atlantic is at present not in a steady state.

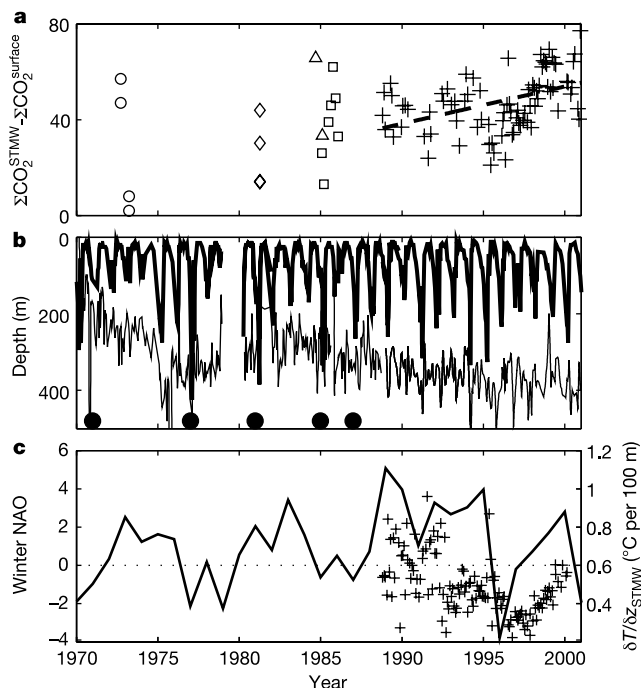


Figure 2 Interannual variability and trends of ΣCO_2 and hydrography in the western North Atlantic near Bermuda (1970–2001). **a**, Difference between surface-layer and STMW ΣCO_2 (in units of $\mu\text{mol kg}^{-1}$) from 1970 to 2001. ΣCO_2 data were sampled at BATS from October 1988 to September 2001 ('plus' symbols). BATS ΣCO_2 data only were adjusted to account for the cyclo-stationary seasonal cycle. Also shown are data from 17 stations near Bermuda collected over the 1972–86 period. ΣCO_2 data from 4 stations were collected in 1972–73 as part of the Geochemical Sections (GEOSECS) program (open circles). ΣCO_2 data from 4 stations were collected in 1981 (open diamonds) as part of the Transient Tracer in the Ocean (TTO) program by T. Takahashi (Lamont Doherty Earth Observatory). ΣCO_2 data from 2 stations were collected in 1984–85 at Hydrostation S by C. D. Keeling (Scripps Institution of Oceanography) (personal communication; triangle). ΣCO_2 data from 7 stations were collected in 1985–86 (squares) at Hydrostation S by P. G. Brewer (Woods Hole Oceanographic Institution)²⁰. Although ΣCO_2 was determined by different methods (potentiometry and manometry), these data sets are comparable to BATS data because we only compare surface and STMW values. The GEOSECS and TTO data sets are available at <http://www.cdiac.ornl.gov>. **b**, Interannual variability and trends of mixed-layer depth (bold line) and depth of the 18°C isotherm (thin line) at BATS from October 1988 to September 2001 and at Hydrostation S from 1970 to 1988. Mixed-layer depth was calculated from CTD profiles and a $0.5 \sigma_\theta$ criterion. STMW was not formed near Bermuda since 1987, during the BATS sampling period. Filled circles denote each year when STMW formed extensively across the subtropical gyre and as far south as Bermuda (32°N). **c**, Interannual variability and trends of the temperature gradient in STMW ($\delta T/\delta z_{\text{STMW}}$) expressed in $^\circ\text{C per 100 m}$, using the methodology of ref. 7 and winter NAO conditions. Climate indices for NAO were compiled by the Climate and Global Dynamics (CGD) division at the National Centre for Atmospheric Research (NCAR; <http://www.cgd.ucar.edu/cas/climind/>). Wintertime (December–March) NAO indices were plotted as a solid line and based on sea-level pressure differences between Lisbon, Portugal and Stykkisholmur, Iceland⁹. $\delta T/\delta z_{\text{STMW}}$ values (plus signs) at BATS were relatively high at ~ 0.2 – $0.8^\circ\text{C per 100 m}$ for the 1988–2001 period, indicating that STMW was not ventilated or formed near Bermuda during the winter. $\delta T/\delta z_{\text{STMW}}$ values below $0.2^\circ\text{C per 100 m}$ are indicative of recent STMW renewal⁷. Dotted line is zero reference line for NAO index.

Table 1 Long-term trends for ΣCO_2 and other parameters at BATS

Parameter	Slope and standard deviation ($\mu\text{mol kg}^{-1} \text{yr}^{-1}$)	Number of observations	r^2
Surface layer (observed trends)			
ΣCO_2	$+1.32 \pm 0.18$	158	0.75
$n\Sigma\text{CO}_2$	$+1.25 \pm 0.14$	158	0.73
DO	-0.42 ± 0.08	155	0.85
Temperature	$+0.009 \pm 0.016$	158	0.94
Salinity	$+0.001 \pm 0.003$	158	0.30
Surface layer (inferred trends)			
dC_{ani}/dt	$+0.9$	NA	NA
dC_{gasex}/dt	$+0.43 \pm 0.43$	158	0.49
dC_{bio}/dt	-0.08 ± 0.09	158	0.22
STMW (observed trends)			
ΣCO_2	$+2.55 \pm 0.25$	96	0.56
$n\Sigma\text{CO}_2$	$+2.64 \pm 0.26$	96	0.55
$\Sigma\text{CO}_2/n\Sigma\text{CO}_2$ ($26.4 \sigma_\theta$)	$+2.22 \pm 0.07$	94	0.65
DO	-0.72 ± 0.21	91	0.17
Nitrate	$+0.07 \pm 0.02$	91	0.14
Phosphate	$+0.004 \pm 0.001$	88	0.15
Temperature	-0.016 ± 0.004	95	0.23
Salinity	-0.002 ± 0.001	96	0.16
STMW (inferred trends)			
dC_{ani}/dt	$+0.9$	NA	NA
dC_{gasex}/dt	$+1.07 \pm 0.25$	88	0.20
dC_{bio}/dt	$+0.55 \pm 0.17$	88	0.25

Long-term trends for ΣCO_2 , $n\Sigma\text{CO}_2$ (that is, ΣCO_2 normalized to a salinity of 36.6), dissolved oxygen (DO), nitrate, phosphate, temperature and salinity are computed for surface-layer and sub-tropical mode water (STMW) at the BATS site from October 1988 to September 2001. Regression statistics (slope, standard deviation, and r^2) were determined using a least-squares fitting routine using a singular value decomposition method (see Methods). Statistical trends were also determined for the biological (C_{bio}) and gas exchange (C_{gasex}) contributions to the observed ΣCO_2 trends (see Methods for details). The trend for the anthropogenic component (C_{ani}) was estimated from the atmospheric CO_2 trend assuming full equilibration between the atmospheric and oceanic perturbation. STMW is defined here by the 17.8 – 18.4°C criterion and typically located at depths from 250 to 400 m. NA, not applicable.

Long-term changes of STMW ΣCO_2 can only result from the uptake of anthropogenic CO_2 from the atmosphere, from changes in the uptake of non-anthropogenic CO_2 from the atmosphere through air–sea gas exchange at the site of STMW formation, or from variations in the remineralization of organic matter along the circulation pathway of the STMW. We evaluate the contributions of these processes to the ΣCO_2 changes by using the long-term changes of inorganic nutrients as indicators of biological changes, by computing the anthropogenic CO_2 contribution from thermodynamic considerations, and by estimating the gas-exchange component by difference. We thus consider the following components:

$$dC/dt = dC_{\text{ant}}/dt + dC_{\text{bio}}/dt + dC_{\text{gasex}}/dt \quad (1)$$

where dC_{ant}/dt is the temporal change in ocean ΣCO_2 as a result of the uptake of anthropogenic CO_2 from the atmosphere, dC_{bio}/dt is the temporal change driven by biological processes such as primary production, remineralization and CaCO_3 formation and dissolution¹⁹, and dC_{gasex}/dt is the temporal change driven by variability in air–sea CO_2 gas exchange as a result of changes in the seawater p_{CO_2} (partial pressure of CO_2) conditions or wind speeds at the site of STMW formation (see Methods for details).

We estimate the rate of change of the anthropogenic component, dC_{ant}/dt , from the atmospheric CO_2 change and the surface ocean buffer factor, assuming that near-surface waters in the subtropical gyres have residence times long enough to equilibrate entirely with the anthropogenic perturbation in atmospheric CO_2 . This assumption is corroborated by long-term observations from various sites as well as many ocean modelling studies¹⁹. We obtain an equilibrium rate of ΣCO_2 increase due to anthropogenic CO_2 of $+0.9 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (Fig. 1c), close to the observed rate of increase in surface ocean $n\Sigma\text{CO}_2$ ($+1.25 \mu\text{mol kg}^{-1} \text{yr}^{-1}$) at BATS. In the surface layer, with the exception of dissolved oxygen (DO), no significant changes in temperature, salinity, or inorganic nutrients were observed from 1988 to 2001 (Table 1). This suggests for the surface layer that long-term changes in the gas-exchange and biological components are small, and that the long-term change in ΣCO_2 is primarily driven by the uptake of anthropogenic CO_2 from the atmosphere.

For STMW, the components that drive the temporal changes in ΣCO_2 are summarized in Table 1 (see Methods and Supplementary Information). This analysis indicates that the rate of $n\Sigma\text{CO}_2$ increase, in excess of equilibration with the anthropogenic perturbation of atmospheric CO_2 , is at least $1.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (that is, $dC_{\text{bio}}/dt + dC_{\text{gasex}}/dt = dC/dt - dC_{\text{ant}}/dt$). As the uptake of anthropogenic CO_2 can only occur at the rate of equilibration or less, this is a lower-bound estimate for the change from gas exchange and biology. Closer inspection reveals that this anomalous increase is primarily due to an increase in the gas exchange component ($dC_{\text{gasex}}/dt = 1.07 \pm 0.25 \mu\text{mol kg}^{-1} \text{yr}^{-1}$) rather than a change in the biological component ($dC_{\text{bio}}/dt = 0.55 \pm 0.17 \mu\text{mol kg}^{-1} \text{yr}^{-1}$). The long-term increase in C_{bio} in STMW is caused by an increase in the inorganic nutrients, and is associated with a long-term decrease in DO. This decrease in DO occurs in a nearly biological stoichiometric ratio to the increase in the inorganic nutrients, corroborating our interpretation (based on nutrient data) that the long-term increase in C_{bio} is due to a slightly enhanced accumulation of remineralized carbon in the STMW, either as a result of increased export or increased residence time of STMW. As there were no significant long-term changes of primary and export production observed over the 1988–2001 period near Bermuda^{15,20}, we view this increase in C_{bio} as an indicator of increased residence time of STMW.

The large changes of the gas exchange component, C_{gasex} (that is, the content of non-anthropogenic atmospheric CO_2) in the STMW layer must have resulted from recent changes in the uptake of CO_2 through gas exchange at the site of STMW formation. Air–sea gas exchange is typically parameterized as a quadratic or cubic function of wind speed²¹ and Δp_{CO_2} (the gradient between air and sea of the

partial pressure of CO_2 , p_{CO_2}). At the site of STMW formation, air-to-sea CO_2 fluxes typically range from 5 to $10 \text{ mmol CO}_2 \text{ m}^{-2} \text{d}^{-1}$ (refs 15, 22) during winter time. If we assume a typical mixed layer thickness of $\sim 400 \text{ m}$ at the site of STMW formation^{4,7}, variability in the air–sea CO_2 flux has the potential to change the ΣCO_2 content of STMW by $\sim 1\text{--}2 \mu\text{mol kg}^{-1}$ each winter. These estimates are consistent with the observed changes in STMW ΣCO_2 and C_{gasex} from 1988 to 2001. There has been a slight increase in mean winter (January–March) daily wind speed ($\sim 0.03 \pm 0.04 \text{ m s}^{-1} \text{yr}^{-1}$; $r^2 = 0.10$) since 1980 in the STMW formation region, potentially equivalent to an increase in STMW ΣCO_2 of $\sim 0.06 \pm 0.08 \mu\text{mol kg}^{-1} \text{yr}^{-1}$. There are insufficient data to evaluate whether seawater p_{CO_2} conditions²² have changed over time at the site of STMW formation (see Fig. 2b legend), but given the large variability in physical conditions at these sites, such changes appear very likely.

The fate of the non-anthropogenic atmospheric CO_2 in STMW (that is, either eventual release to the atmosphere or transport to the ocean interior) may be linked to interannual variability of STMW formation and changes in NAO state. Extensive STMW formation, as far south as Bermuda, occurred during numerous winters in the 1960s and 1970s, and lastly in 1985 and 1987 (Fig. 2b, c), typically coinciding with NAO negative phases^{4,7,10}. However, since 1987, winter STMW formation has been limited, coinciding with a predominantly NAO positive phase^{4,7,10}. Winter mixing near Bermuda has been generally weak ($<200 \text{ m}$ deep), and not deep enough to entrain ΣCO_2 from the STMW layer into the mixed layer at BATS (Fig. 2b). Analyses of the vertical temperature gradients within STMW⁷ (that is, $\delta T/\delta z_{\text{STMW}}$) also indicate that the STMW has not formed near Bermuda and that the STMW layer at BATS has not been ventilated recently (Fig. 2c), in agreement with our finding of a small increase in the C_{bio} component in the absence of a change in biological export. During this period of relatively weak winter mixing, atmospheric CO_2 is retained and trapped in the STMW layer. As STMW is transported to the western boundary current, water parcels cool and sink to deeper depths in ‘cooling spirals’^{23,24}, and CO_2 is thereby transferred to deeper depths, constituting a potential long-term sink of CO_2 ($>10 \text{ yr}$).

In contrast, during periods of extensive deep winter mixing ($>300\text{--}350 \text{ m}$ deep, before 1988, and primarily associated with NAO negative phases^{4,7,10}), mixing entrains water from the STMW layer and ‘liberates’ atmospheric CO_2 from STMW. We estimate that at least two-thirds of the STMW CO_2 accumulation would be redistributed to seasonal thermocline waters overlying the STMW layer (see Methods). If this model is correct, atmospheric CO_2 absorbed during STMW formation is only stored for short periods in STMW, thereby constituting a short-term sink of CO_2 ($<3 \text{ yr}$). The relatively short observation period makes it difficult to demonstrate the proposed links between STMW variability, NAO changes and fate of CO_2 within STMW. This is further complicated by the time delay of about 3–4 yr (ref. 11) between the time of STMW formation and the arrival of this signal near Bermuda.

The increase in C_{gasex} implies a significant and recent change in the role of mode waters as an oceanic sink for atmospheric CO_2 . Given current estimates about the volume of the North Atlantic STMW ($1,672 \times 10^3 \text{ km}^3$; ref. 10), rates of STMW formation (5 to 23 Sv ; $1 \text{ Sv} = 10^6 \text{ m}^3 \text{s}^{-1}$) and residence times^{5,25–29}, and the CO_2 increase (that is, $dC_{\text{gasex}}/dt + dC_{\text{bio}}/dt = 1.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$), we estimate that $\sim 0.4\text{--}2.8 \text{ Pg C}$ (or $0.02\text{--}0.24 \text{ Pg C yr}^{-1}$) has accumulated in the subtropical gyre of the North Atlantic over the past 13 yr (see Methods and Supplementary Information). This accumulation of atmospheric CO_2 in STMW, in excess of the rate of equilibration of anthropogenic CO_2 , represents an extra oceanic sink of CO_2 that is 3–10% of the average global uptake rate of CO_2 ($\sim 2 \text{ Pg C yr}^{-1}$)³ for the 1988–2001 period. In earlier decades, such as the 1970s and 1980s, extensive STMW ventilation of STMW across the gyre should have regularly ‘liberated’ a considerable amount of CO_2 trapped in the STMW layer, thereby reducing the oceanic uptake of

atmospheric CO₂ into STMW. The present transfer of CO₂ via STMW to the ocean interior, facilitated by enhanced uptake of CO₂ and weak winter mixing, should continue until extensive deep winter mixing and STMW ventilation returns to the subtropical gyre. Interannual variability in the uptake, storage and transfer of STMW CO₂ to the ocean interior provides another factor and feedback controlling the global ocean uptake of CO₂. □

Methods

Attribution of ΣCO₂ changes

We use the concept of quasi-conservative tracers¹⁹ to separate the contribution of gas exchange, anthropogenic CO₂ and biological processes to the observed change in ΣCO₂ over time (dC/dt). We estimate the biological contribution to dC/dt from the temporal changes in the phosphate concentration (PO₄^{3−}), nitrate concentration (NO₃[−]) and the changes in total alkalinity (TA), assuming constant stoichiometric ratios between carbon and phosphate during photosynthesis and respiration ($r_{C:P} = 117:1$)¹⁹. This gives $dC_{bio}/dt = r_{C:P} dPO_4^{3-}/dt - 1/2(dTA/dt + dNO_3^-/dt)$, where all concentrations are salinity normalized and given in μmol kg^{−1}. We estimate the anthropogenic change dC_{ant}/dt by assuming that surface waters in the subtropical gyre of the North Atlantic have followed the atmospheric perturbation and therefore have taken up anthropogenic CO₂ in equilibrium with anthropogenic CO₂ in the atmosphere. We compute dC_{ant}/dt using theoretical considerations of the oceanic carbonate system and the 1.31 μatm yr^{−1} increase in atmospheric CO₂ observed at the island of Bermuda between 1988 and 2001 (data collected by NOAA Climate Monitoring and Diagnostics Laboratory at the Bermuda west station, and available at <http://www.cmdl.noaa.gov>). This yields a rate of increase of about 0.9 μmol kg^{−1} yr^{−1}. Finally, we estimate the change in ΣCO₂ as a result of changes in gas exchange from the difference of the observed increase and the biological and anthropogenic changes, that is, $dC_{gasex}/dt = dC/dt - dC_{ant}/dt - dC_{bio}/dt$. As this last term is computed by difference, it contains all the accumulated errors made in the computation of dC_{bio}/dt and dC_{ant}/dt . However, as it turns out, the dC_{bio}/dt term is relatively small, and the uncertainty in the estimated equilibrium anthropogenic CO₂ change is small. Furthermore, the true anthropogenic CO₂ change can only be equal or smaller than the estimated equilibrium change, and therefore our estimated dC_{gasex}/dt term has to be regarded as a lower-bound estimate. If we allow for 20% uncertainty in both of our dC_{bio}/dt and dC_{ant}/dt estimates, our gas exchange term will be uncertain by about 20% as well.

Estimate of long-term trends

We compute the long-term trends of the data by fitting the data to a function that includes a constant, a linear trend and harmonic terms with periods of 12, 6 and 4 months. The 8 unknown parameters of this fit are determined by minimizing the misfit between this function and the data in a least squares sense using singular value decomposition. In Table 1, only the parameter of the linear trend is given. The uncertainty of the parameters has been estimated from the square root of the variance of the fit multiplied with the square root of $\chi^2/(N - 2)$, where χ^2 is the minimized misfit and N is the number of data points.

Estimate of STMW redistribution during winter deep mixing

Some portion of CO₂ within the STMW layer will be entrained into the seasonal thermocline during winters of deep mixing (>250–300 m). STMW typically has a thickness of 100–150 m, and is located at a depth of 250–400 m throughout the subtropical gyre, whereas the lower depth limit of the seasonal thermocline is typically 200–250 m. For example, if a 100-m-thick layer of STMW is entrained into a 300-m mixed layer during winter mixing, only one-third of the original STMW CO₂ will be retained within the STMW layer upon subsequent seasonal stratification (that is, 200 m deep). This redistribution should decrease or dilute ΣCO₂ contents in the STMW layer and increase ΣCO₂ contents in the seasonal thermocline.

Estimate of oceanic sink of CO₂ in STMW layer since 1988

We estimate the oceanic sink of CO₂ from empirical and model understanding of the physical structure and circulation of the STMW layer in the subtropical gyre, and a simple input–output steady-state budget. We assume that the volume of the STMW is constant, which implies that input (STMW formation rate) equals output (subduction of STMW to deeper layers of the ocean). We furthermore assume that the observed rate of increase at BATs is representative for the entire STMW layer in the North Atlantic. The total accumulation of carbon in STMW (ΔC_{Inv}), where 'Inv' is the inventory of carbon can then be estimated by multiplying the 13-yr ΣCO₂ increase ($\Delta \Sigma CO_2 = dC/dt'$, where t' is 13 yr) with the volume of STMW (V_{STMW}). As the volume of STMW is somewhat uncertain, we can alternatively determine it from estimated STMW formation rates (ϕ_{STMW}) and residence times (τ). In summary, this gives: $\Delta Inv = \Delta \Sigma CO_2 V_{STMW} = \Delta \Sigma CO_2 \tau \phi_{STMW}$.

We estimate the volume of STMW between the permanent and seasonal thermocline in the North Atlantic using volume data computed for 0.1 σ_θ density layers by G. Daniels and D. Olson (Univ. Miami), with density derived from the temperature and salinity climatology of the World Ocean Atlas (http://www.nodc.noaa.gov/OC5/data_woa.html). This yields volumes of STMW in the North Atlantic of 1.66×10^{15} m³, 1.79×10^{15} m³, and 2.54×10^{15} m³ for the 26.4–26.6, 26.4–26.7 and 26.4–26.8 σ_θ layers, respectively. The 26.4–26.6 σ_θ density criterion compares well with previous estimates of the volume of STMW of 1.67×10^{15} m³ (ref. 10). Lower (0.54 Pg C) and upper (2.23 Pg C) bound estimates of the CO₂ accumulation in STMW are determined on the basis of this range of volume estimates. For example, given the three different volumes of the STMW layer and a rate of nΣCO₂

increase of $1.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (that is, $dC_{gasex}/dt + dC_{bio}/dt$), we estimate the accumulation of CO₂ of 0.54, 0.58 and 0.90 Pg C, respectively, over the 1988–2001 period. A rate of nΣCO₂ increase of $2.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (that is, $dC_{ant}/dt + dC_{gasex}/dt + dC_{bio}/dt$) yields an accumulation of CO₂ of 0.88, 0.94 and 1.34 Pg C, respectively over the 1988–2001 period. In this scheme, we assume a mean residence time (τ) of 10 yr for water in the STMW layer. After STMW formation, the average time for a parcel of STMW to be transported from the site of formation to the western boundary current along the path of gyre recirculation has been reported as ranging from 6 to 14 yr (refs 11, 28, 29). ³H/³He tracer data indicate that the typical age of water in STMW observed near Bermuda is 3–4 yr old¹¹, with the age of STMW increasing along the mean geostrophic circulation pathway from the site of STMW formation to the western boundary current^{24,25}. If τ is shorter (that is, 6 yr), this would yield an accumulation of CO₂ of 1.47, 1.57 and 2.23 Pg C, respectively over the 1988–2001 period.

In the second approach, we use current knowledge about the annual rate of STMW formation and residence time of water in the STMW layer to estimate the total volume of the STMW layer. The annual supply of new STMW (ϕ_{STMW}) from the site of STMW formation has been estimated through empirical and model studies to range from 5 to 23 Sv (Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$)^{5,25–27}. As above, we assume a mean residence time (τ) of 10 yr for water in the STMW layer. Lower (0.32 Pg C) and upper (2.8 Pg C) bound estimates of the CO₂ accumulation in STMW are estimated using this method. For example, a rate of nΣCO₂ increase of $1.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (that is, $dC_{gasex}/dt + dC_{bio}/dt$) at three STMW formation rates (5, 14 and 23 Sv) yields an accumulation of CO₂ of 0.32, 0.90 and 1.47 Pg C, respectively over the 1988–2001 period. A rate of nΣCO₂ increase of $2.6 \mu\text{mol kg}^{-1} \text{yr}^{-1}$ (that is, $dC_{ant}/dt + dC_{gasex}/dt + dC_{bio}/dt$) at the 3 STMW formation rates (5, 14 and 23 Sv) yields an accumulation of CO₂ of 0.61, 1.72 and 2.83 Pg C, respectively, over the 1988–2001 period.

Received 27 December 2001; accepted 17 October 2002; doi:10.1038/nature01253.

- Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
- Hanawa, K. & Talley, L. D. *Ocean Circulation and Climate* International Geophysics Series (eds Siedler, G., Church, J. & Gould, J.) Vol. 77, 378–386 (Academic, San Diego, 2001).
- Prentice, C. et al. *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. et al.) 183–237 (Cambridge Univ. Press, Cambridge, 2001).
- Klein, B. & Hogg, N. On the interannual variability of 18 degree water formation as observed from moored instruments at 55°W. *Deep-Sea Res.* **43**, 1777–1806 (1996).
- Hazeleger, W. & Drijfhout, S. S. Mode water variability in a model of the subtropical gyre: response to anomalous forcing. *J. Phys. Oceanogr.* **28**, 266–288 (1998).
- Joyce, T. M., Deser, C. & Spall, M. A. The relation between decadal variability of subtropical mode water and the North Atlantic Oscillation. *J. Clim.* **13**, 2550–2569 (2000).
- Alfuitis, M. A. & Cornillon, P. Annual and interannual changes in the North Atlantic STMW layer properties. *J. Phys. Oceanogr.* **31**, 2066–2086 (2001).
- Rodwell, M. J., Rowell, D. P. & Folland, K. K. Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Nature* **398**, 320–323 (1999).
- Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
- Worthington, L. V. The 18° water in the Sargasso Sea. *Deep-Sea Res.* **5**, 297–305 (1959).
- Jenkins, W. J. Studying subtropical thermocline ventilation and circulation using tritium and ³He. *J. Geophys. Res.* **103**, 15817–15831 (1998).
- Talley, L. & Raymer, M. Eighteen degree water variability. *J. Mar. Res.* **40** (suppl.), 757–775 (1982).
- Jenkins, W. G. On the climate of the subtropical gyre: Decade timescale variation in water mass renewal in the Sargasso Sea. *J. Mar. Res.* **40** (suppl.), 265–290 (1982).
- Joyce, T. M. & Robbins, P. The long-term hydrographic record at Bermuda. *J. Clim.* **9**, 3121–3131 (1996).
- Bates, N. R. Interannual changes of oceanic CO₂ and biogeochemical properties in the Western North Atlantic subtropical gyre. *Deep-Sea Res.* **48**, 1507–1528 (2001).
- Bates, N. R., Michaels, A. F. & Knap, A. H. Seasonal and interannual variability of the oceanic carbon dioxide system at the U.S. JGOFS Bermuda Atlantic Time-series Site. *Deep-Sea Res.* **43**, 347–383 (1996).
- Keeling, C. D. *The Global Carbon Cycle* (ed. Heimann, M.) 413–430 (Springer, New York, 1993).
- Karl, D. M. et al. Temporal studies of biogeochemical processes in the world's oceans during the JGOFS era. *Deep-Sea Res.* (in the press).
- Gruber, N. & Sarmiento, J. L. *The Sea: Biological-Physical Interactions in the Ocean* (eds Robinson, A. R., McCarthy, J. J. & Rothschild, B. J.) Vol. 12, 337–399 (Wiley and Sons, New York, 2002).
- Conte, M. H., Ralph, N. & Ross, E. H. Seasonal and interannual variability in deep ocean particle fluxes at the Ocean Flux Program (OFP)/Bermuda Atlantic Time-series Study (BATs) site in the western Sargasso Sea near Bermuda. *Deep-Sea Res.* **48**, 1471–1507 (2001).
- Wanninkhof, R. & McGillis, W. R. A cubic relationship between air-sea CO₂ exchange and wind speed. *Geophys. Res. Lett.* **26**, 1889–1892 (1999).
- Takahashi, T. et al. Global air-sea flux of CO₂ based on surface ocean pCO₂, and seasonal biological and temperature effects. *Deep-Sea Res.* **49**, 1601–1622 (2002).
- Behringer, D. & Stommel, H. The beta-spiral in the North Atlantic subtropical gyre. *Deep-Sea Res.* **A** **27**, 225–238 (1980).
- Spall, M. A. Cooling spirals and recirculation in the subtropical gyre. *J. Phys. Oceanogr.* **22**, 564–571 (1992).
- Woods, J. D. & Barkmann, W. A lagrangian mixed layer model of Atlantic 18° water formation. *Nature* **319**, 574–576 (1986).
- Speer, K. & Zippman, E. Rates of water mass formation in the North Atlantic Ocean. *J. Phys. Oceanogr.* **22**, 93–104 (1992).
- Marsh, R. & New, A. L. Modeling 18° water variability. *J. Phys. Oceanogr.* **26**, 1059–1080 (1996).
- Luyten, J. R., Pedlosky, J. & Stommel, H. The ventilated thermocline. *J. Phys. Oceanogr.* **13**, 292–309 (1983).
- Williams, R. G., Spall, M. A. & Marshall, J. C. Does Stommel's mixed layer "demon" work. *J. Phys. Oceanogr.* **25**, 3089–3102 (1995).

30. Graedel, T. E., Brewer, P. G. & Rowland, F. S. Panel 4. Chemistry of the air-sea interface. *Appl. Geochem.* 3, 37–48 (1988).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank A. H. Knap, A. F. Michaels, D. A. Hansell, D. K. Steinberg and C. A. Carlson for their contributions to the BATS program. We also thank the numerous technicians, graduate students, and captains and crew of the RV *Weatherbird II*, who have contributed to the success of the BATS program since 1988. W. S. Broecker, T. Takahashi, C. D. Keeling and P. G. Brewer are thanked for the use of ΣCO_2 data collected in the Sargasso Sea during 1972–86. This work was supported by the National Science Foundation and National Oceanographic and Atmospheric Administration.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.B. (e-mail: nick@bbsr.edu).

The role of volatiles in magma chamber dynamics

Herbert E. Huppert & Andrew W. Woods

Institute of Theoretical Geophysics, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW, UK; and the BP Institute of Multiphase Flow, Madingley Rise, Madingley Road, Cambridge CB3 0EZ, UK

Many andesitic volcanoes exhibit effusive eruption activity¹, with magma volumes as large as 10^7 – 10^9 m^3 erupted at rates of 1 – $10 \text{ m}^3 \text{ s}^{-1}$ over periods of years or decades. During such eruptions, many complex cycles in eruption rates have been observed, with periods ranging from hours to years^{2–7}. Longer-term trends have also been observed, and are thought to be associated with the continuing recharge of magma from deep in the crust and with waning of overpressure in the magma reservoir. Here we present a model which incorporates effects due to compressibility of gas in magma. We show that the eruption duration and volume of erupted magma may increase by up to two orders of magnitude if the stored internal energy associated with dissolved volatiles can be released into the magma chamber. This mechanism would be favoured in shallow chambers or volatile-rich magmas and the cooling of magma by country rock may enhance this release of energy, leading to substantial increases in eruption rate and duration.

Consider a magma with bulk density ρ in a chamber of volume V undergoing a mass recharge rate Q_i and eruption rate Q_0 , as sketched in Fig. 1. Conservation of mass indicates that:

$$\frac{d}{dt}(\rho V) = \rho \frac{dV}{dt} + V \frac{d\rho}{dt} = Q \quad (1)$$

where $Q = Q_i - Q_0$. The density of the magma, which consists of melt, crystals and gas, can be written, in general form, as $\rho = \rho[p, T, x(T), N]$ where p is pressure, T temperature, $x(T)$ the mass fraction of crystals in the magma and N the total mass fraction of volatiles. Depending on the pressure, a fraction of these volatiles are dissolved in the magma and the remainder exist in the gaseous phase. Differentiating this expression for the density and incorporating the result into equation (1), we obtain:

$$\frac{dV}{dt} + \frac{V \partial \rho}{\rho \partial p} \frac{dp}{dt} = \frac{Q}{\rho} - \frac{V \partial \rho}{\rho \partial T} \frac{dT}{dt} \quad (2)$$

where the second term on the right-hand side represents the rate of change of volume associated with the change in density with

temperature of the magma, which includes the effects of the change of crystal content with temperature, leading to an effective thermal expansion.

It is known^{8–10} that the rate of change in volume of the chamber, dV/dt , is related to the associated rate of change in pressure, dp/dt , as a result of deformation of the surrounding rock by:

$$\frac{1}{\beta_r} \frac{dp}{dt} = \frac{1}{V} \frac{dV}{dt} \quad (3)$$

where β_r is the effective bulk modulus of the surrounding wall rock. The exact value of β_r depends on the density of microfractures in the rock, but a value of 10^{10} Pa is typical^{10,11}. Incorporating equation (3) into equation (2), we obtain:

$$V \left[\frac{1}{\beta_r} + \frac{1}{\rho} \frac{\partial \rho}{\partial p} \right] \frac{dp}{dt} = \frac{Q}{\rho} - \frac{V \partial \rho}{\rho \partial T} \frac{dT}{dt} \quad (4)$$

Thus, as expressed by equation (3), the effective bulk modulus of the compressible magma, β , is given by:

$$\frac{1}{\beta} = \frac{1}{\beta_r} + \frac{1}{\rho} \frac{\partial \rho}{\partial p} \quad (5)$$

We show here that for typical volatile-rich magmas, the second term on the right-hand side of equation (5) is much larger than the first term, which indicates that the compressibility of the saturated magma plus exsolved gas, C , defined as the inverse of the effective bulk modulus β , greatly exceeds that of the wall rock.

An explicit relationship for ρ in terms of the gas density ρ_g , the crystal density ρ_c and the melt density ρ_m is:

$$\rho = \left[\frac{n}{\rho_g} + (1-n) \left(\frac{x}{\rho_c} + \frac{1-x}{\rho_m} \right) \right]^{-1} \quad (6)$$

where, to a good approximation, the gas density follows the ideal gas law¹², $\rho_g = p/RT$, where R is the universal gas constant. We assume the exsolution of gas occurs at thermodynamic and chemical equilibrium, which is valid for the slow processes considered here^{13,14}. Thus, in our model, the exsolved volatile content, n , follows from Henry's law^{12,13,15}. This law, for water, which is frequently the dominant volatile species present, is $n = N - sp^{1/2}(1-x) \geq 0$, on the assumption that the magma is saturated, where $s = 4 \times 10^{-6} \text{ Pa}^{-1/2}$ for water vapour. This will occur at sufficiently low pressures and sufficiently high total volatile or crystal contents. Alternatively, if $sp^{1/2}(1-x) \geq N$, the magma is undersaturated and $n \equiv 0$. Analogous calculations can be made for

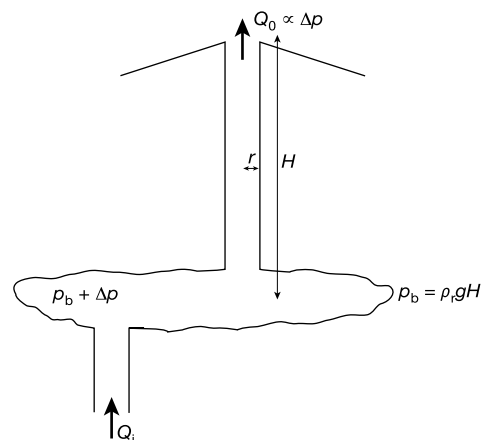


Figure 1 Schematic of magma reservoir. In the reservoir there is an overpressure Δp to a background pressure field at depth H given by $p_b = \rho_r g H$, where ρ_r is the density of the rocks above the magma chamber and g is the acceleration due to gravity. During a slow effusive eruption, through a conduit of radius r , mass is input at rate Q_i and erupted at rate Q_0 .

magmas in which the dominant volatile content is carbon dioxide. In all the calculations which follow here the crystals and melt are assumed to have bulk modulus $\beta_c = 10^{10}$ Pa (refs 10, 11).

Figure 2a and b illustrates the variation of effective compressibility $C = 1/\beta$ with pressure for different volatile and crystal contents. As shown in Fig. 2, the depth may be related to the pressure through the lithostatic gradient. For deep, crystal-poor magmas the magma is unsaturated and hence the bulk modulus is of

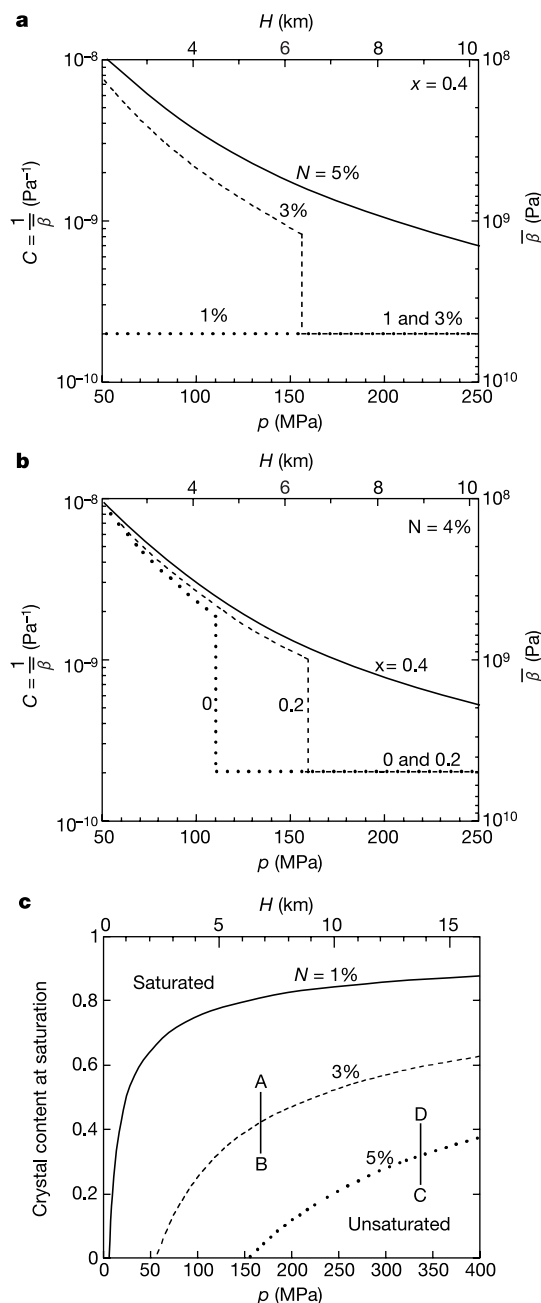


Figure 2 Magma compressibility as a function of pressure. Data are shown for various values of magma volatile content with $x = 0.4$ (a) and magma crystal content with $N = 4\%$ (b). The discontinuity in magma compressibility occurs when the magma is just saturated. When it is undersaturated there is no exsolved gas and the compressibility is the sum of the compressibility of the wall rock and that of the magma contained within the chamber. c, Critical crystal content as a function of pressure for various values of volatile content. In these calculations $s = 4 \times 10^{-6} \text{ Pa}^{-1/2}$; $\rho_c = 2,600 \text{ kg m}^{-3}$; $\rho_r = 2,500 \text{ kg m}^{-3}$; $\rho_m = 2,300 \text{ kg m}^{-3}$; $R = 462 \text{ J K}^{-1} \text{ kg}^{-1}$ and $T = 800^\circ \text{C}$ (as appropriate for silicic magma).

order 10^{10} Pa and independent of depth. For shallower systems, the magma is saturated and the much larger compressibility of the accompanying gas phase leads to a bulk modulus of order 10^8 – 10^9 Pa. The transition in C occurs when the magma first becomes saturated. At this critical point, a small decrease in pressure leads to expansion of the mixture primarily through exsolution of relatively low-density gas, whereas a small increase of pressure can only be accommodated through compression of the liquid magma and the country rock. Figure 2c illustrates the crystallinity at which the magma first becomes saturated as a function of pressure or depth for various volatile contents. For larger crystal contents, gas is exsolved and the compressibility greatly increases (Fig. 2a, b). The greater compressibility leads to eruption of a larger mass to relieve the same overpressure and hence the duration of the resulting effusive eruption will be longer.

To evaluate the detailed response during an eruption, equations (4) and (6), being heavily nonlinear, need to be solved numerically. However, the typical change in pressure during an eruption is only of order 1–10 MPa, and Fig. 2a and b illustrates that at depths of 3–10 km, such changes in pressure lead to small variations in compressibility, typically $\leq 10\%$, except just at the saturation point (see below). As an effective approximation, we may therefore take the chamber compressibility and the chamber volume as being constant, at the mean value, during an eruption.

For given magma and flow conduit properties, the eruption rate is a function of the chamber overpressure^{16–18}. As a simple illustrative calculation, we follow ref. 19 and assume that, for a magma chamber at depth H ,

$$Q_0 = (\rho S A^2 / H \mu) \Delta p \equiv \lambda \Delta p \quad (7)$$

where S is a shape factor of about 0.1, A the conduit cross-section and μ the effective viscosity of the magma in the conduit. At

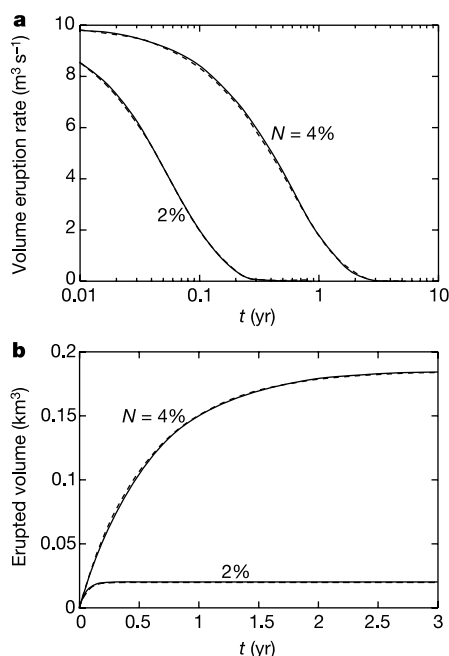


Figure 3 Eruption volume and volume eruption rate. a, Volume eruption rate Q_0 , which is proportional to the chamber overpressure Δp , as expressed in equation (7). b, Erupted volume V_e , as functions of time for magma at a depth, H , of 5 km. The calculations incorporate a crystal content $x = 0.4$ and volatile contents of 2 wt% (a typical undersaturated magma) and 4 wt% (a typical saturated magma). The conduit radius is 15 m, and dynamic viscosity $\mu = 10^7$ Pa. The chamber volume is 10 km^3 and initial chamber overpressure is 10 MPa. Solid lines show predictions of the approximate model equations (9) and (10) with $\beta = 127.6 \text{ MPa}$, and dashed lines are numerical solutions incorporating the full variation of β .

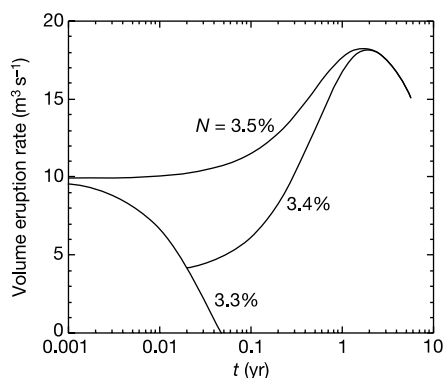


Figure 4 Variation of eruption rate for a silicic magma at a depth of 5 km with 3.5, 3.4 and 3.3 wt% volatiles, accounting for the cooling of the magma at a rate of 10^{-6} K s^{-1} according to equation (4). After less than one tenth of a year, the magma with a 3.4 wt% volatile content becomes saturated and hence much more compressible; also the chamber pressure begins to increase owing to the increasing volume of gas. This leads to an increase in the eruption rate and a much longer eruption duration and erupted mass. Magma with 3.3 wt% volatile content remains unsaturated, and has a much more rapidly waning eruption rate, while magma with 3.5 wt% volatiles is already saturated and shows a gradually increasing eruption rate with time owing to the crystallization and the associated gas release.

constant temperature, the chamber overpressure, Δp , then evolves as:

$$\frac{d\Delta p}{dt} = \{Q_i - \lambda \Delta p\}(\bar{\beta}/\rho V) \quad (8)$$

which has a solution for an eruption initiated at an overpressure Δp_0 at $t = 0$ of

$$\Delta p = \Delta p_0 \exp\left(-\frac{\lambda \bar{\beta} t}{\rho V}\right) + \left(\frac{\bar{\beta}}{\rho V}\right) \int_0^t \exp\left[-\frac{\lambda \bar{\beta}}{\rho V}(t-s)\right] Q(s) ds \quad (9)$$

As an example, for constant Q_i , equation (9) reduces to:

$$\Delta p = [\Delta p_0 - Q_i/\lambda] \exp[-\lambda \bar{\beta} t/\rho V] + (Q_i/\lambda) \quad (10)$$

whereas the volume erupted, V_e , is given by:

$$V_e = \frac{V}{\bar{\beta}} (\Delta p_0 - Q_i/\lambda) \left[1 - \exp\left(-\frac{\lambda \bar{\beta} t}{\rho V}\right)\right] + Q_i t/\rho \quad (11)$$

Figure 3 shows predictions of equations (10) and (11) for two typical eruptions (with $Q_i = 0$). This illustrates how the presence of volatiles in the magma reservoir leads to an increase in eruption time and erupted volume by a factor of 30 for the particular chosen conditions. We note that the timescale of the eruption $\tau = \rho V/\lambda \bar{\beta}$ and that the total erupted volume $V_e = V \Delta p_0/\bar{\beta}$. Both of these increase with increasing exsolved volatile content, through the factor $1/\bar{\beta}$, as well as with increasing chamber volume V . Figure 3 also illustrates the accuracy of the approximation by the very close agreement with a full numerical solution (dashed line).

Our model points to the crucial role that exsolved volatiles in the magma chamber play in controlling eruption volume and duration during effusive eruptions in which the eruption rate is controlled by the overpressure. As the effective compressibility of the magma increases, a given decrease in pressure leads to a greater expansion of the magma and therefore, for a given chamber size, a greater mass of magma is erupted. In turn, the density of the remaining gas–magma mixture is smaller, which leads to a smaller residual mass in the chamber. In addition, during such eruptions magma heating or cooling may also occur^{9,19,20}, especially after input of hot basaltic magma at the base of the chamber¹⁸. This brings in the last term on the right-hand side of equation (4). Heating and cooling lead to changes in magma crystallinity and in turn this can change the volatile saturation in the chamber (for example, $A \rightarrow B$, $C \rightarrow D$ in

Fig. 2c). Such changes in volatile saturation may greatly change the eruption rate. Figure 4 illustrates the increase in eruption rate that may occur as the magma is cooled against the country rock¹⁰ so that it eventually becomes saturated in volatiles owing to the additional crystallization. At this stage the compressibility increases by an order of magnitude (Fig. 2), the chamber pressure also increases owing to the exsolution and the timescale of the continuing eruption increases by a factor of ten.

Although our model is idealized, transitions in volatile saturation within a magma chamber may help to provide new explanations for the abrupt declines in eruption regimes, such as observed during dome eruptions at Mt Redoubt, Alaska²¹, in 1992, Mt Unzen, Japan⁵, in 1995 and at Soufrière Hills, Montserrat⁷, in 1998. In reality, flows in conduits are more complex than modelled here and may have multiple flow states^{16,17,22,23}. Our simple linear relationship (7) captures the behaviour on one solution branch and points to the dominant control of volatiles in the chamber on eruption longevity. Transitions from one steady state to a different solution branch can lead to more complex eruption histories and this needs to be considered alongside our analysis of phenomena that occur only in the magma chamber. The challenge now is to collect and interpret field data with the appropriate theoretical models in mind. $\square$

Received 13 August; accepted 14 October 2002; doi:10.1038/nature01211.

1. Sigurdsson, H. (ed.) *Encyclopedia of Volcanoes* (Academic, San Diego, 2000).
2. Rose, W. I. Pattern and mechanism of volcanic activity at the Santiaguito volcanic dome, Guatemala. *Bull. Volcanol.* **36**, 73–94 (1972).
3. Rose, W. I. Volcanic activity at the Santiaguito volcano, 1976–1984. *Geol. Soc. Am.* **212**, 17–27 (1987).
4. Swanson, D. A. & Holcomb, R. T. *Lava Flows and Domes; Emplacement and Hazard Implications* (ed. Fink, J. H.) 3–24 (Springer, Berlin, 1990).
5. Nakada, S., Shimizu, H. & Ohta, K. Overview of the 1990–1995 eruption at Unzen Volcano. *J. Volcanol. Geotherm. Res.* **89**, 1–22 (1999).
6. Belousov, A., Belusova, B. & Voight, B. Multiple edifice failures, debris avalanches and associated eruptions in the Holocene history of Shiveluch volcano, Kamchatka, Russia. *Bull. Volcanol.* **61**, 324–342 (1999).
7. Sparks, R. S. J. & Young, S. R. *The Eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999* (eds Druitt, T. H. & Kokelaar, B. P.) 45–69 (Memoir, Geological Society, London, 2002).
8. Blake, S. Volcanism and the dynamics of open magma chambers. *Nature* **289**, 783–785 (1981).
9. Blake, S. Volatile oversaturation during the evolution of silicic magma chambers as an eruption trigger. *J. Geophys. Res.* **89**, 8237–8244 (1984).
10. Tait, S., Jaupart, C. & Vergnolle, S. Pressure, gas content and eruption periodicity of a shallow, crystallising magma chamber. *Earth Planet. Sci. Lett.* **92**, 107–123 (1989).
11. Touloukian, Y. S., Judd, W. R. & Roy, R. F. *Physical Properties of Rocks and Minerals* Vol. 1 (McGraw Hill, New York, 1981).
12. Huppert, H. E., Turner, J. S. & Sparks, R. S. J. The effects of volatiles on mixing in calcalkaline magma systems. *Nature* **297**, 554–557 (1982).
13. Sparks, R. S. J. The dynamics of bubble formation and growth in magmas: A review and analysis. *J. Volcanol. Geotherm. Res.* **3**, 1–37 (1978).
14. Hurwitz, S. & Navon, O. Bubble nucleation in rhyolitic melts: Experiments at high pressure, temperature and water content. *Earth Planet. Sci. Lett.* **122**, 267–280 (1994).
15. Burnham, C. W. & Jahns, R. H. A method for determining the solubility of water in silicate melts. *Am. J. Sci.* **260**, 721–745 (1962).
16. Melnik, O. E. & Sparks, R. S. J. Non-linear dynamics of lava dome extrusion. *Nature* **402**, 37–41 (1999).
17. Barmin, A., Melnik, O. & Sparks, R. S. J. Periodic behaviour in lava dome eruptions. *Earth Planet. Sci. Lett.* **199**, 173–184 (2002).
18. Couch, S., Sparks, R. S. J. & Carroll, M. R. Mineral disequilibrium in lavas explained by convective self-mixing in open magma chambers. *Nature* **411**, 1037–1039 (2001).
19. Stasiuk, M., Jaupart, C. & Sparks, R. S. J. On the variations of flow rate in non-explosive lava eruptions. *Earth Planet. Sci. Lett.* **114**, 505–516 (1993).
20. Flock, A. & Marti, J. The generation of overpressure in felsic magma chambers by replenishment. *Earth Planet. Sci. Lett.* **163**, 301–314 (1997).
21. Miller, T. P. & Chouet, B. A. The 1989–1990 eruptions of Redoubt volcano—an introduction. *J. Volcanol. Geotherm. Res.* **62**, 1–10 (1994).
22. Jaupart, C. Gas loss from magmas through conduit walls during eruption. *Geol. Soc.* **289**, 783–785 (1998).
23. Massol, H. & Jaupart, C. The generation of gas overpressure in volcanic eruptions. *Earth Planet. Sci. Lett.* **166**, 57–70 (1999).

Acknowledgements We thank S. Sparks for comments and suggestions, and S. Blake and C. Jaupart for critiques on a first draft of this paper.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.E.H. (e-mail: heh1@esc.cam.ac.uk).

The role of parasites in sympatric and allopatric host diversification

Angus Buckling* & Paul B. Rainey†‡

* Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

† Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

‡ School of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand

Exploiters (parasites and predators) are thought to play a significant role in diversification, and ultimately speciation, of their hosts or prey^{1–3}. Exploiters may drive sympatric (within-population) diversification if there are a variety of exploiter-resistance strategies or fitness costs associated with exploiter resistance^{4–8}. Exploiters may also drive allopatric (between-population) diversification by creating different selection pressures and increasing the rate of random divergence^{9,10}. We examined the effect of a virulent viral parasite (phage) on the diversification of the bacterium *Pseudomonas fluorescens* in spatially structured microcosms¹¹. Here we show that in the absence of phages, bacteria rapidly diversified into spatial niche specialists with similar patterns of diversity across replicate populations. In the presence of phages, sympatric diversity was greatly reduced, as a result of phage-imposed reductions in host density decreasing competition for resources. In contrast, allopatric diversity was greatly increased as a result of phage-imposed selection for resistance, which caused populations to follow divergent evolutionary trajectories. These results show that exploiters can drive diversification between populations, but may inhibit diversification within populations by opposing diversifying selection that arises from resource competition.

Evidence for the role of exploiters in diversification of natural populations is limited. Indirect evidence supports a role for exploiters in sympatric phenotypic diversification, for example, ornamentation in molluscs¹² and body size in sticklebacks^{13,14}, whereas genetic diversity in traits associated with parasite resistance between geographic regions^{15,16} suggest a role for exploiters in allopatric diversification. Uncertainty over the precise role of exploiters stems from difficulties associated with the study of populations in their natural environments: long generation times and the inability to control for environmental and genetic variation. These problems can be overcome using laboratory populations of bacteria. The large size of these populations (approximately 10⁹ cells), combined with short generation times, favours rapid evolution. Moreover, ability to precisely control both environment and genetic composition of founding cells allows identification of causal mechanisms¹⁷. Here, we carried out an experiment to address the role of exploiters in driving both sympatric and allopatric diversification of experimental bacterial populations. Because resource competition has been shown to be important for sympatric diversification^{11,18,19}, our experiments were carried out under conditions where ecological diversification through resource competition also operated.

Experimental populations of the plant-colonizing bacterium *Pseudomonas fluorescens* SBW25 (ref. 20) provide a useful system for studying the evolution of diversity^{11,21,22}. When propagated in spatially structured environments (a static glass microcosm containing nutrient-rich medium), isogenic *P. fluorescens* populations rapidly diversify, generating numerous niche specialist types that are readily distinguished by their (heritable) colony morphologies on agar plates. These can be grouped into three distinct classes on the basis of colony morphology and niche occupation. Smooth morphotypes, resembling the ancestral morphotype, inhabit the liquid phase; wrinkly spreader morphotypes form a biofilm at the air–

broth interface; and fuzzy spreader morphotypes colonize the harsher, less aerobic bottom of the vials¹¹. Competition for resources is responsible for the origin and maintenance of diversity, as demonstrated by the operation of negative frequency-dependent selection^{11,21}; this is a fitness advantage when types are rare because there is less intense competition within their niche^{23,24}.

To examine the effect of an exploiter on diversification we introduced a naturally virulent phage SBW25φ2 (ref. 25) into the spatially structured microcosms. Virulent phages are viruses with life cycles that result in bacterial death (invasion of bacterial cells, replication and cell lysis), and hence impose strong selection for bacterial resistance²⁶. We inoculated twenty-four microcosms with isogenic *P. fluorescens* SBW25. Twelve were simultaneously inoculated with isogenic SBW25φ2. Populations were independently propagated by transferring 1% of the culture to a fresh microcosm every two days. Bacterial densities and diversities were determined at each transfer.

In apparent contrast to much theory addressing the role of exploiters in driving sympatric diversification^{5–8}, diversity within microcosms was significantly lower in populations evolving with phages than without (Fig. 1a; using mean diversity through time: $t = 12.01$, degrees of freedom d.f. = 14, $P < 0.001$). As previously observed¹¹, populations evolving in the absence of phages rapidly diversified into a variety of niche specialist morphotypes, resulting in high levels of sympatric diversity throughout the experiment

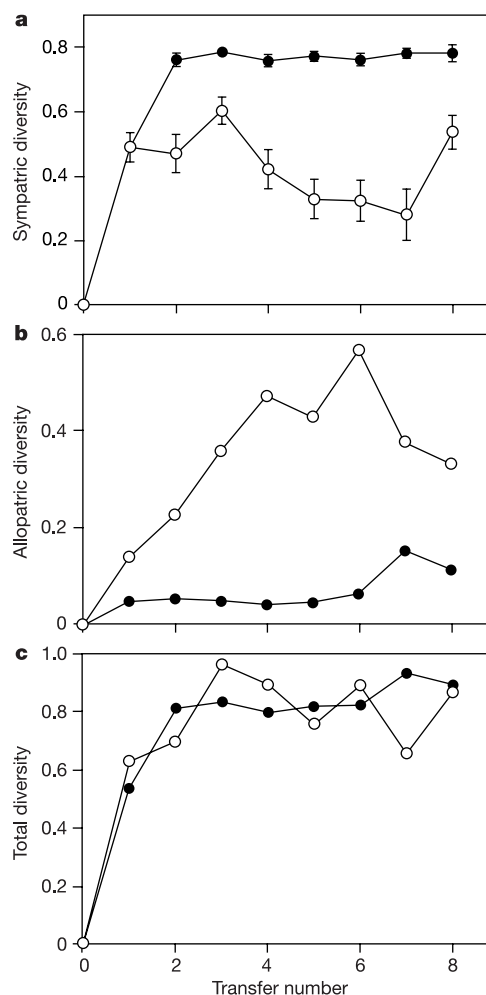


Figure 1 Partitioning of diversity. Mean (± 1 s.e.m., $n = 12$) sympatric bacterial diversity (a), allopatric diversity (b), and total diversity (c) through time for populations evolving in isolation (filled symbols) and evolving with phages (open symbols).

(Fig. 1a). Grouped smooth morphotypes and wrinkly spreader morphotypes (up to five of each within each population) coexisted at approximately equal frequencies, whereas fuzzy spreader morphs generally made up less than 1% of the populations. Populations evolving with phages also diversified, but to a much lesser extent. These populations consisted of fewer morphotypes (at detectable frequencies) and were dominated by single morphotypes. Recognizing that diversity of colony morphology may not reflect high levels of diversity in resistance, we checked resistance traits with colony types in assays where colonies were challenged by ancestral and contemporary phage populations. No evidence of any variation in resistance traits within morphotypes was found.

Allopatric diversity was much higher in populations evolving with phages than without (Fig. 1b; sign test of number of times allopatric diversity was greater in populations with phages than without; 8/8, $P < 0.01$). Populations evolving without phages contained many of the same morphotypes at similar frequencies, whereas populations evolving with phages tended to be dominated by different morphotypes in different populations. In two out of twelve of these populations we observed high frequencies of bacteria that produced highly mucoid colonies that colonized the sides of the microcosm. These were not observed in any of the phage-free

populations, suggesting phage-imposed selection was in part responsible for the evolution of this new form. We noticed that total diversity—the sum of sympatric and allopatric diversities²⁷—did not differ between populations evolving in the presence or absence of phages (Fig. 1c; sign test: $P > 0.2$). Instead, phage presence greatly altered the partitioning of sympatric and allopatric diversity. However, total diversity measured across all populations in the experiment (combining populations with and without phages) was consistently greater than total diversity of populations evolving either with or without phages (sign test: $P < 0.01$ in both cases), which suggests that the presence of exploiters in some, but not all, populations may increase diversity over a geographic scale.

What is the explanation for phage-imposed reductions in sympatric diversity and simultaneous increase in allopatric diversity? Selection for resistance to phages is likely to have played a critical role: bacteria evolved from being entirely sensitive to entirely resistant to the ancestral phages within two transfers, and the frequency of resistance remained at greater than 0.95 in all populations (data not shown). Furthermore, these populations of bacteria and phages underwent antagonistic coevolution: continual cycles of bacteria evolving resistance to phages and phages evolving infectivity to these 'resistant' bacteria²⁵. Coevolution was driven largely by directional selection, favouring bacterial resistance and phage infectivity to an increasingly wide range of phage and bacterial populations, respectively²⁵. Selection for greater resistance ranges imposed by the continual evolution of infective phages is likely to oppose the negative-frequency dependent (diversifying) selection maintaining the diversity of spatial niche specialists, decreasing sympatric diversity. Because all genetic variation was freshly generated by mutation in this study, beneficial resistance mutations are likely to have appeared in different genetic backgrounds in different populations. Selection would therefore have favoured different morphotypes in different populations, increasing allopatric diversity.

However, diversities might also have been affected by phage-imposed reductions in bacterial density: phages caused an average fourfold reduction in host density, relative to phage-free populations, in the first experiment ($t = 10.13$, d.f. = 15, $P < 0.0001$). Reductions in host density may reduce resource competition between niche specialists, reducing the strength of frequency-dependent selection²⁴, and therefore sympatric diversity. A reduction in the strength of frequency-dependent selection may also increase the role of chance in determining which morphotypes dominated in different populations, hence increasing allopatric diversity.

To clarify the roles of selection for resistance and phage-imposed density reductions, we carried out an additional experiment. We chose four bacterial populations that had evolved with phages and differed in resistance to their respective phage populations. Six replicate populations of each of the four original populations were grown for two transfers under three different conditions: in the presence of their respective phage populations, in the absence of phages and in the absence of phages but at a low bacterial density so as to mimic the density-reducing effect of phages. Sympatric and allopatric diversity was independently regressed onto the final bacterial density, combining the phage-free treatments (control and low density) together. The slope of sympatric diversity and density was not affected by phage (Fig. 2a, c; treatment by density interaction: $F_{1,8} = 3.1$, $P > 0.1$), suggesting phage-imposed density reductions were largely responsible for low sympatric diversity in the presence of phage. Indeed, density was able to explain 70% of the total deviance in sympatric diversity ($F_{1,9} = 80.2$, $P < 0.0001$). In contrast, population density did not have a significant effect on allopatric diversity: there was no relationship between density and allopatric diversity in phage-free populations (Fig. 2b, c; changing slope to zero caused a non significant increase in deviance, $F_{1,8} = 0.01$, $P > 0.2$).

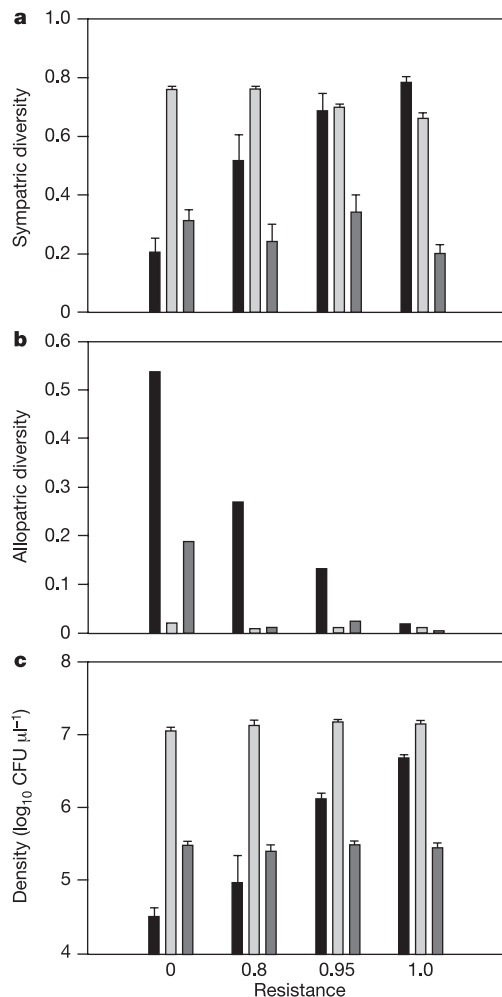


Figure 2 Clarifying the roles of selection for resistance and density. Mean (± 1 s.e.m., $n = 6$) sympatric diversity (**a**), allopatric diversity (**b**), and mean (± 1 s.e.m., $n = 6$) density ($\log_{10}$ colony forming units, CFU; **c**) of populations of bacteria grown for two transfers (15 generations). Populations varied in the proportion of bacteria resistant to their respective future phage populations (x axis). Populations were subject to three treatments: with phages (black bars), phage-free control (light grey bars) and phage-free, low density (dark grey bars).

We wondered whether selection for resistance was able to explain the increased allopatric diversification in the presence of phages, as phage-imposed reductions in host density were unable to do so. Under both hypotheses, (phage-imposed selection for resistance and density reductions) the intensity of selection for resistance (as determined by the proportion of sensitive cells in the population) should determine the impact of phages on the subsequent diversity of bacterial populations. Thus, population resistance is expected to correlate positively with sympatric diversity and negatively with allopatric diversity in the presence of phages; no such relationship between resistance and diversity is expected in phage-free populations. As predicted, the effect of resistance on both sympatric and allopatric diversity differed between populations with and without phages (Fig. 2a, b; resistance by treatment interaction: $F_{1,4} = 15.48$, $P < 0.02$; $F_{1,4} = 19.65$, $P = 0.01$, respectively), with no relationships in phage-free populations (Fig. 2a, b; $P > 0.2$, in both cases) and a positive (Fig. 2a; $t = 9.14$, d.f. = 4, $P < 0.001$) and a negative (Fig. 2b; $t = 9.2$, d.f. = 4, $P < 0.001$) relationship, respectively, in populations with phages. The negative relationship (Fig. 2b) between resistance and allopatric diversity in populations with phages must therefore be largely a direct result of selection for resistance rather than phage-imposed density reductions. Resistance mutations occurred in different genetic backgrounds in different populations, causing populations to follow divergent evolutionary trajectories. Persistent antagonistic coevolution would have further fuelled allopatric diversification by preventing the subsequent evolution of morphotypes that consistently have a selective advantage in the absence of phage-imposed selection.

We cautiously suggest that these results are likely to have general relevance to diversification of host and prey species. Here we show that parasites increase diversification between populations through a combination of selection for resistance and chance: the morphotype in which the resistance mutation occurred (which varied between populations) swept to high frequencies and determined subsequent population evolutionary trajectories. Chance is likely to be equally important in determining the relative success of newly evolved host or prey types when there is strong selection for exploiter avoidance. Furthermore, antagonistic coevolution, or the continual introduction of new exploiter species, is likely to further fuel this diversification by continually imposing selection for new defence traits. That exploiters might have decreased the degree of diversification within geographical regions is more contentious: much theory examining diversification predicts the opposite result^{5–8}. However, unlike some ecological theory^{28,29}, theory addressing diversification does not assume the simultaneous operation of resource competition and exploiter-mediated apparent competition. Both types of competition are likely to be important in diversification within geographical regions, but may have opposite effects: parasite-induced mortality may relax resource competition. This may reduce the selective advantage of newly arisen types capable of exploiting alternative resources, and hence diversification. □

Methods

Culturing techniques

Populations were propagated in static 25 ml glass universal bottles (microcosms) containing 6 ml King's medium B (KB) in a static 28 °C incubator. Twenty-four replicate microcosms were inoculated with 10^8 cells of *Pseudomonas fluorescens* isolate SBW25 (grown for 24 h at 28 °C in an orbital shaker at 2.86g) and 10^5 clonal particles of a naturally associated DNA phage SBW25Φ2 were added to twelve. 60 µl of each culture was transferred to fresh medium every 2 days, for eight transfers (approximately 60 bacterial generations). Cultures were frozen at –80 °C in 20% glycerol at every second transfer. Bacterial densities were estimated by plating diluted cultures onto standard KB agar, and counting the number of colony forming units (CFUs) after 48 h of incubation at 28 °C.

Measurement of diversity

Diversity was measured by determining the morphologies of 100 random colonies on KB agar plates^{11,21,22}. Sympatric diversity was calculated as the complement of Simpson's index

of concentration $(1 - \lambda)$ (ref. 30).

$$1 - \lambda = \left(1 - \sum_i p_i^2\right) \left(\frac{N}{N-1}\right) \quad (1)$$

where p_i is the proportion of the i th morph and N is the total number of colonies sampled. This measure is the probability that two randomly selected morphotypes are different. Allopatric diversity was calculated as a variance (d^2) (ref. 27): the additional probability that two types, randomly chosen from any population, are different:

$$d^2 = \sum_i (p_{ij} - \bar{p}_i)^2 \quad (2)$$

where $\bar{p}_i$ is the mean proportion of the i th morph across all j populations.

Resistance assays

Ten independent colonies of each morphologically distinct variant within each population were streaked across a line of phages on KB agar (see ref. 25). Phages were isolated by centrifuging cultures with 10% chloroform, which lysed and pelleted bacterial debris. Bacteria were assayed against ancestral phage and contemporary phage populations. Bacteria were scored as sensitive if there was evidence of growth inhibition in the presence of phages, otherwise they were scored as resistant.

Identifying the mechanisms

We isolated four populations of bacteria from the second transfer that differed in resistance (between 0 and 1 proportion of resistant bacteria) to their respective phage populations from two transfers in the future. Future phage populations rather than contemporary phage populations were used because resistance to contemporary phage was generally very high. Resistance was determined as above from 100 bacterial colonies. Phages were removed from these populations by dilution, and the resulting cultures grown for 24 h at 28 °C, 2.86g. Three experimental treatments, each consisting of six replicate microcosms, were established from each of the four populations. Microcosms in the first treatment were established with 10^7 bacterial cells and 10^5 phage particles from their respective future phage populations. The second treatment was a control: 10^7 bacterial cells but no phages were added. The third treatment addressed the impact of phage-independent density reductions on diversity, with microcosms initiated with 10^3 cells. After 2 days, a portion of the culture, either 10^7 cells, or 10^3 cells in the low density treatment, was transferred to fresh medium and grown for a further 24 h. Cultures were plated onto KB agar and diversity measured as above. Data were analysed as Generalized Linear Models using GLIM4. The relative importance of selection for resistance and density on diversity was determined by fitting treatment as a factor (combining control and low density treatments) and ($\log_{10}$) density as a covariate. The effect of phage on mean sympatric and allopatric diversity was determined by fitting treatment as a factor (excluding low density treatment) and resistance as a covariate; the significance of slopes was determined by t -tests.

Received 19 July; accepted 24 September 2002; doi:10.1038/nature01164.

1. Darwin, C. *The Origin of Species* (Thompson and Thomas, Chicago, 1872).
2. Lack, D. *Darwin's Finches* (Cambridge Univ. Press, Cambridge, 1947).
3. Schluter, D. Ecological character displacement in adaptive radiation. *Am. Nat.* **157** (suppl.), S4–S16 (2000).
4. Holt, R. D. Predation, apparent competition, and the structure of prey communities. *Theor. Pop. Biol.* **12**, 197–229 (1977).
5. Brown, J. S. & Vincent, T. L. Organisation of predator-prey communities as an evolutionary game. *Evolution* **46**, 1269–1283 (1992).
6. Schluter, D. *The Ecology of Adaptive Radiations* (Oxford Univ. Press, Oxford, 2000).
7. Abrams, P. A. Character shifts of prey species that share predators. *Am. Nat.* **156** (suppl.), S46–S61 (2000).
8. Doebeli, M. & Dieckmann, U. Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *Am. Nat.* **156** (suppl.), S77–S101 (2000).
9. Travis, J. The significance of geographical variation in species interactions. *Am. Nat.* **148** (suppl.), S1–S8 (1996).
10. Thompson, J. N. Specific hypotheses on the geographic mosaic of coevolution. *Am. Nat.* **153** (suppl.), S1–S14 (1999).
11. Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394**, 69–72 (1998).
12. Stone, H. M. I. On predator deterrence by pronounced shell ornament in epifaunal bivalves. *Paleontology* **41**, 1051–1068 (1998).
13. Walker, J. A. Ecological morphology of lacustrine threespine stickleback *Gasterosteus aculeatus* L. (*Gasterosteidae*) body shape. *Biol. J. Linn. Soc.* **61**, 3–50 (1997).
14. Vamori, S. M. & Schluter, D. Impacts of trout predation on the fitness of sympatric sticklebacks and their hybrids. *Proc. R. Soc. Lond. B* **269**, 923–930 (2002).
15. Berenbaum, M. R. & Zangerl, A. R. Chemical phenotype matching between a plant and its insect herbivore. *Proc. Natl Acad. Sci. USA* **95**, 13743–13748 (1998).
16. Kaltz, O. & Shykoff, J. Local adaptation in host-parasite systems. *Heredity* **81**, 361–370 (1998).
17. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. 1. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).
18. Helling, R. B., Vargas, C. N. & Adams, J. Evolution of *Escherichia coli* during growth in a constant environment. *Genetics* **116**, 349–358 (1987).
19. Schluter, D. Experimental evidence that competition promotes divergence in adaptive radiation. *Science* **266**, 798–801 (1994).
20. Rainey, P. B. & Bailey, M. J. Physical and genetic map of the *Pseudomonas fluorescens* SBW25 chromosome. *Mol. Microbiol.* **19**, 521–533 (1996).
21. Buckling, A., Kassen, R., Bell, G. & Rainey, P. B. Disturbance and diversity in experimental microcosms. *Nature* **408**, 961–964 (2000).

22. Kassen, R., Buckling, A., Bell, G. & Rainey, P. B. Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature* **406**, 508–512 (2000).
23. Ayala, F. J. & Campbell, C. A. Frequency-dependent selection. *Annu. Rev. Ecol. Syst.* **5**, 115–138 (1974).
24. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).
25. Buckling, A. & Rainey, P. B. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. Lond. B* **269**, 931–936 (2002).
26. Lenski, R. E. & Levin, B. R. Constraints on the coevolution of bacteria and virulent phage—a model, some experiments, and predictions for natural communities. *Am. Nat.* **125**, 585–602 (1985).
27. Lande, R. Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76**, 5–13 (1996).
28. Holt, R. D., Grover, J. & Tilman, D. Simple rules for interspecific dominance in systems with exploitative and apparent competition. *Am. Nat.* **144**, 741–771 (1994).
29. Leibold, M. A graphical model of keystone predators in food webs: trophic regulation of abundance, incidence and diversity patterns in communities. *Am. Nat.* **147**, 784–812 (1996).
30. Simpson, E. H. Measurement of diversity. *Nature* **163**, 688 (1949).

Acknowledgements We thank M. Brockhurst, L. Hurst and S. West for comments on the manuscript. This work was supported by NERC (UK) and the Royal Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B. (e-mail: bssagib@bath.ac.uk).

Calcium activation of BK_{Ca} potassium channels lacking the calcium bowl and RCK domains

Rebecca Piskorowski & Richard W. Aldrich

Department of Molecular and Cellular Physiology, and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305-5345, USA

In many physiological systems such as neurotransmitter release, smooth muscle relaxation and frequency tuning of auditory hair cells, large-conductance calcium-activated potassium (BK_{Ca}) channels create a connection between calcium signalling pathways and membrane excitability^{1–4}. BK_{Ca} channels are activated by voltage and by micromolar concentrations of intracellular calcium. Although it is possible to open BK_{Ca} channels in the absence of calcium^{5–9}, calcium binding is essential for their activation under physiological conditions. In the presence of intracellular calcium, BK_{Ca} channels open at more negative membrane potentials^{5,10–14}. Many experiments investigating the molecular mechanism of calcium activation of the BK_{Ca} channel have focused on the large intracellular carboxy terminus, and much evidence supports the hypothesis that calcium-binding sites are located in this region of the channel. Here we show that BK_{Ca} channels that lack the whole intracellular C terminus retain wild-type calcium sensitivity. These results show that the intracellular C terminus, including the ‘calcium bowl’ and the RCK domain, is not necessary for the calcium-activated opening of these channels.

Several experimental results imply that the intracellular C terminus of the BK_{Ca} channel mediates Ca²⁺ binding and/or activation. The channel’s pore-forming subunit is shown in Fig. 1a. Mutations in a conserved group of aspartic acid residues, called the ‘calcium bowl’, change the ability of low concentrations of Ca²⁺ to activate the channel¹⁵. Additional experiments indicate that the last 400 residues of the C terminus (which includes the calcium bowl) are important in Ca²⁺ activation of the BK_{Ca} channel¹⁶. Removal of these amino acids yields a non-functional BK_{Ca} channel, and co-expression of the 400 amino acid fragment with a shortened channel restores normal behaviour^{16,17}. Replacement of this region by the

corresponding region of a BK_{Ca} channel homologue that lacks Ca²⁺ sensitivity (Slo3) results in a channel with greatly decreased Ca²⁺ sensitivity¹⁶. In addition, this region of the C terminus binds Ca²⁺ in ⁴⁵Ca²⁺-overlay protein blot assays^{18,19}, and combinations of point mutations in the calcium bowl and RCK domains eliminate Ca²⁺ activation^{20,21}. The structure of a bacterial K⁺ channel gated by millimolar Ca²⁺ has been solved in the presence of 200 mM Ca²⁺ (ref. 22), and Ca²⁺-binding sites are clearly evident in the C-terminal RCK domain. Taken together, these results have led to the commonly accepted conclusion that the calcium bowl and the RCK domain, either separately or together, contain Ca²⁺-binding sites that are involved in activation of the channel by micromolar concentrations of Ca²⁺. If this idea is correct, then removing these domains should greatly alter the Ca²⁺ activation of the channel.

Counter to this prediction, we found that a truncated channel that lacks the whole intracellular C terminus (made by deleting the amino acids after residue 323; Fig. 1a) retains Ca²⁺ sensitivity (Fig. 2). Single-channel current traces (Figs 1b and 2a) and quantified voltage- and Ca²⁺-dependent activity (Fig. 2b, c) of wild-type and truncated channels showed that the truncated channel is activated over the same range of Ca²⁺ concentrations as the full-length channel. Thus, a pore-forming domain with an intact C-terminal domain is not required for voltage- and Ca²⁺-dependent gating. Given the variability in gating between the different channels (Fig. 2c), it was unclear whether the apparent change in voltage dependence (Fig. 2b) is a genuine property of the truncated channel. The behaviour of the truncated and full-length channels was not identical: single-channel kinetics showed that the truncated channel has a shorter mean open time (Fig. 3), indicating that the transition energy of the open state of the channel has been altered.

Expression of the truncated channel, as determined by electrophysiological recordings, was roughly a thousand-fold less than that

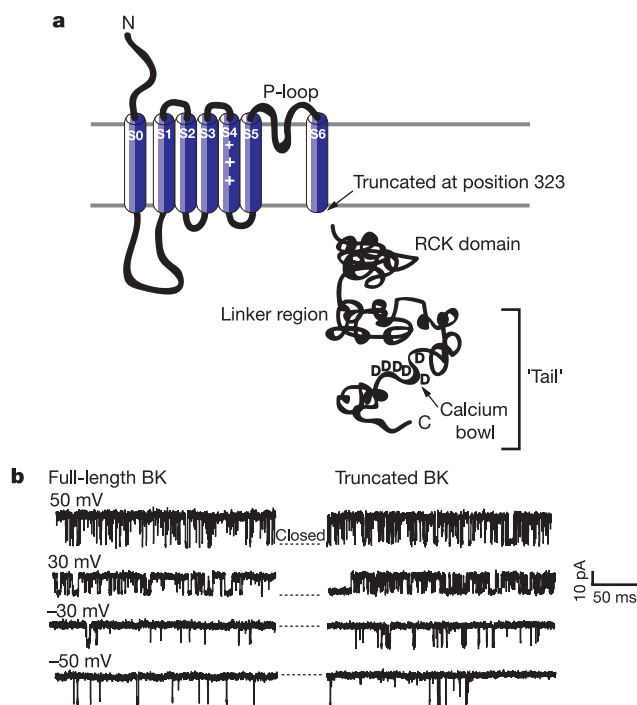


Figure 1 The truncated BK_{Ca} channel retains normal function. **a**, The truncated BK_{Ca} channel. The different domains of the intracellular C terminus and the aspartic acids in the calcium bowl are indicated. The ‘tail’ corresponds to the last 400 amino acids of the C terminus that have been proposed to be important in Ca²⁺ activation of BK_{Ca} channels. **b**, Single-channel currents measured at 100 μM Ca²⁺ from full-length and truncated channels heterologously expressed in *Xenopus* oocytes.

of the wild-type channel in *Xenopus* oocytes and HEK 293 cells. This reduced surface expression of the truncated channel was probably due to disrupted trafficking. A very similar truncated form of BK_{Ca} channel is retained in the endoplasmic reticulum in MDCK cells²³. Immunohistochemical assessment of COS cells transfected with a truncated BK_{Ca} channel fused to an amino-terminal c-Myc tag showed that the truncated channel was localized predominantly in the endoplasmic reticulum (data not shown). This is consistent with the previously described role of the C-terminal domain in trafficking and tetramerization²⁴.

Endogenous BK_{Ca} channels are expressed in very small amounts in *X. laevis* oocytes²⁵. Thus, the low expression of the heterologously expressed BK_{Ca} channel raised the possibility that our recordings might be from endogenous channels or from heteromultimers of introduced and endogenous channel subunits. The expression levels of the heterologously expressed channels were, however, much greater than those of endogenous *Xenopus* oocyte BK_{Ca} channels, for which one channel in about 5% of patches from uninjected cells was previously observed²⁵. We typically observed 2–5 channels per patch but rarely more than 7. Given a roughly 5% incidence of endogenous channels in patches, the probability of observing two channels in a patch would be 0.0025, five channels in a patch 3×10^{-7} , and seven channels in a patch 7.8×10^{-10} .

To verify that the single-channel currents being measured did originate from the truncated channel, we generated the point mutation Tyr294Val in the truncated BK_{Ca} channel construct. Endogenous and truncated BK_{Ca} channels show a roughly 75% reduction in current in the presence of 1 mM external tetraethylammonium (TEA; ref. 25 and Fig. 4c), whereas homomeric Tyr294Val channels are resistant to external TEA block^{26,27}. Likewise, heteromultimers of wild-type and Tyr294Val mutant BK_{Ca} subunits show a graded external TEA block, depending on the number of Tyr294Val subunits that are in present the channel^{26,27}. Because external TEA causes a fast block, it is possible to estimate the number of Tyr294Val subunits in a channel by observing the effect

of TEA on the apparent single-channel current amplitude when the records are low-pass-filtered at 1 kHz (refs 26, 27).

If the truncated Tyr294Val BK_{Ca} channel subunits formed heterotetramers with endogenous oocyte channel subunits, then a reduction in the apparent single-channel current would be observed in the presence 3–6 mM external TEA. At millimolar concentrations of TEA, homomeric Tyr294Val channels did show a slight reduction of apparent single-channel current when the data were filtered at a low bandwidth; however, this slight block (5–10%) was clearly discernible from the block observed when a single wild-type subunit was present in the tetramer (25–30%).

Incorporation of the Tyr294Val mutation in the truncated BK_{Ca} channel rendered the channel resistant to the external TEA block, as seen from the small change in single-channel amplitude (Fig. 4a, b). The different dose–response curves for the truncated BK_{Ca} channel with and without the Tyr294Val mutation are shown in Fig. 4c. In eight outside-out patches containing 18 BK_{Ca} channels, the presence of 3–6 mM external TEA blocked the channel only slightly (5–10%). We also recorded currents from 22 inside-out patches from oocytes injected with messenger RNA of the truncated Tyr294Val BK_{Ca} channel with 3, 6 or 10 mM TEA in the patch pipet. The Ca²⁺ sensitivity and single-channel conductance of these channels were indistinguishable from those of truncated Tyr294Val BK_{Ca} channels in the absence of TEA. These data confirm that small numbers of truncated channel subunits reach the plasma membrane and form functional homomeric Ca²⁺-gated channels. Our results are not from endogenous channels or from heteromultimeric channels containing combinations of introduced truncated and endogenous wild-type subunits. Ca²⁺-sensitive truncated channels were also observed in HEK 293 cells (Fig. 4d), which lack endogenous Ca²⁺-activated currents²⁸.

The most straightforward interpretation of our data is that the Ca²⁺-activation site is unchanged in the truncated BK_{Ca} channel and is therefore not located in the C terminus, the calcium bowl or the RCK domain. This means that the mechanism of Ca²⁺-dependent gating of BK_{Ca} channels at micromolar (physiological) concentrations of Ca²⁺ is different from the mechanism proposed for the millimolar Ca²⁺-dependent gating in the bacterial K⁺ channel MthK²². Given that the truncated channel has near normal conductance, voltage-dependent gating and Ca²⁺-dependent gating, the proposed mechanism of MthK channel opening, through a torsion on the S6 domain owing to Ca²⁺ binding to the RCK domain²², does not seem to apply to BK_{Ca} channels. A few previous mutagenesis studies have suggested candidate Ca²⁺-binding sites in regions outside the C terminus¹⁹.

Another possible explanation of our results is that the Ca²⁺ sensitivity of BK_{Ca} channels is mediated by an unknown accessory

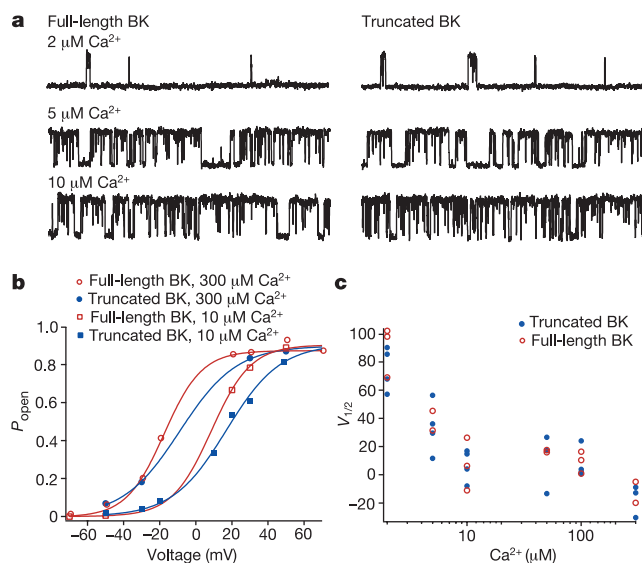


Figure 2 The truncated BK_{Ca} channel is activated by micromolar levels of Ca²⁺. **a**, Single-channel traces at 30 mV and three different Ca²⁺ concentrations. **b**, Boltzmann-fitted plots of P_{open} as a function of voltage for the full-length and truncated channels. **c**, Plot of half activation voltage $V_{1/2}$, as a function of internal Ca²⁺ concentration. $V_{1/2}$ values were derived from fits similar to those shown in **b**. For data recorded at 2 μM Ca²⁺ the maximum value of P_{open} was estimated to be the same as that observed at higher Ca²⁺ concentrations. Filled and open circles represent measurements from several patches expressing truncated or full-length channels, respectively.

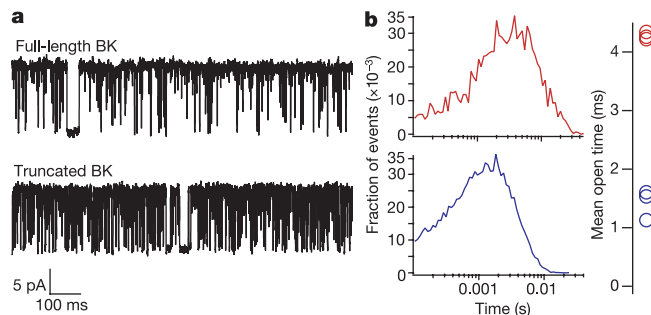


Figure 3 Mean open time is altered in truncated BK_{Ca} channels. **a**, Single-channel traces at 50 mV and 100 μM Ca²⁺ for the full-length and truncated BK_{Ca} channel. **b**, Histogram of the open lifetime for the full-length and truncated channel, and plot of the mean open time for each.

protein. The low expression of the truncated channel in both *Xenopus* oocytes and HEK 293 cells leaves open the possibility that a soluble protein can replace the intracellular C terminus and function as a Ca^{2+} -binding domain for the truncated channel. If this is true, then the factor or factors would have to be present in both expression systems used in these experiments, that is, in *Xenopus* oocytes and HEK 293 cells. A tight association between this accessory protein and the truncated BK_{Ca} channel would be required for Ca^{2+} activation to be seen in cell-free patches.

The ability of a BK_{Ca} channel without a cytoplasmic C-terminal domain to be gated by micromolar concentrations of Ca^{2+} seems incompatible with previous work implicating the calcium bowl and the RCK domain in Ca^{2+} binding. But locating the binding site of a ligand, or determining its properties, is very difficult when the binding is inferred from channel activation, particularly if the binding is energetically coupled to a conformational change²⁹. Most studies implicating the C-terminal domain in Ca^{2+} binding were based on functional measurements, from which binding properties are inferred from the ability of Ca^{2+} to modulate the

voltage range of channel activation. The few direct Ca^{2+} -binding measurements that have been reported were carried out on isolated C-terminal domains under conditions in which it is impossible to relate binding to the activation state of the channel. In a complex, coupled system such as BK_{Ca} channel gating there are many ways in which to alter or disrupt the Ca^{2+} sensitivity of channel opening without directly affecting the Ca^{2+} -binding site itself. Because Ca^{2+} binding, channel opening and voltage sensing are all allosterically coupled^{5,7,12–14,30}, mutations that alter the ability of Ca^{2+} to activate the channel might not affect the Ca^{2+} -binding site directly; these mutations might instead disrupt the allosteric coupling between Ca^{2+} binding, voltage sensing and/or channel opening. Our present understanding of BK_{Ca} channel function indicates that simpler models describing the relationship between voltage, Ca^{2+} concentrations and open state stability are inadequate to interpret the mechanisms of individual mutations^{5,7,12–14,30}.

Our experiments are not prone to the problems of interpreting mutant channel behaviour from allosteric coupling between Ca^{2+} binding and channel gating. Because micromolar concentrations of Ca^{2+} can activate BK_{Ca} channels after the intracellular C terminus has been removed, it can be concluded that the removed region is not essential in the Ca^{2+} activation of the protein. Retention of Ca^{2+} sensitivity in the absence of the C-terminal cytoplasmic domain implies that the calcium bowl and the RCK domain do not make up the 'high-affinity' Ca^{2+} -binding sites in BK_{Ca} channels. Instead, the intracellular C terminus may function as a domain that mediates or modulates the allosteric coupling between Ca^{2+} binding and opening conformational changes, as suggested by the changes in the single-channel kinetics of the truncated channel. □

Methods

Mutagenesis

The wild-type construct used in these experiments was the *mbr5* clone of the mouse homologue of the *mslo* gene, which was a gift from L. Salkoff. We made the truncated mutant by inserting a stop codon at position Ile 323 (C-terminal sequence: VPEI-stop) using the mutagenic primer: 5'-GTTTCCTGCGCGCGCTTAGATTTCAGGGAC-3' and the Expanded High Fidelity PCR System (Roche). The 2.4 kilobases of wild-type coding sequence that remained after the inserted stop codon were removed from the vector. We sequenced the resulting construct repeatedly with several different sequencing primers to ensure that no changes had been made to the gene besides the deletion. The complementary DNA of all constructs was propagated in a modified Blue Script vector BS-MXT (Stratagene). This vector was transformed into One Shot TOP10 Competent cells (Invitrogen). Antisense RNA (cRNA) was transcribed from both vectors with T3 polymerase from the mMessage mMachine kit (Ambion). We injected 0.05–5 ng of cRNA into *X. laevis* oocytes 1–3 d before recording. For the mammalian expression system, the truncated and full-length *mslo1* coding regions were subcloned into the pcDNA3.1 hygro plasmid. About 1 μg of DNA was transfected onto plates using Lipofectamine 2000 (Invitrogen).

Electrophysiology

Excised patches in inside-out and outside-out configurations were transferred into a separate chamber and washed with at least 50 volumes of internal solution. All experiments were done at 22°C. Patches were allowed to stabilize for at least 5–10 min before recording. Internal solutions contained (in mM): 140 KMeSO₃, 6 KCl, 20 HEPES buffer and 40 μM (+)-18-crown-6-tetracarboxylic acid (18C6TA), pH 7.2. CaCl₂ was buffered by HEDTA or nitrilotriacetic acid to give the indicated free Ca^{2+} concentrations. Final Ca^{2+} concentrations were measured with a Ca^{2+} electrode. The external solution contained (in mM): 140 KMeSO₃, 2 KCl, 2 MgCl₂ and 20 HEPES buffer, pH 7.2. Where appropriate, TEA was added to the external solutions and the pH adjusted. Patch electrodes were made with borosilicate glass (VWR Micropipettes). The tips of the electrodes were coated with sticky wax (Kerr Corp) and fire-polished just before use. Pipette access resistance (0.8–2 M Ω) was measured in the bath solution. Data were acquired using an Axopatch 200-A patch clamp amplifier (Axon Instruments) in the resistive feedback mode. An Apple Macintosh-based computer system using Pulse acquisition software (HEKA Elektronik) and an ITC-16 hardware interface (Instrutech Scientific Instruments) were used to collect data. All records were digitized at 20- μs intervals and low-pass-filtered at 8 kHz. Further digital filtering was accomplished with software as needed. We analysed data by TAC 4.02 software (Bxuxton Corporation), the ScanApp analysis program (developed in-house by D. Perkins and D. Cox) and IgorPro (Wavemetrics). Patches containing more than one channel were analysed by fitting amplitude histograms with sums of gaussian functions. The P_{open} (open probability) measurements for these patches were derived from these fits and also by integrating the digitized record.

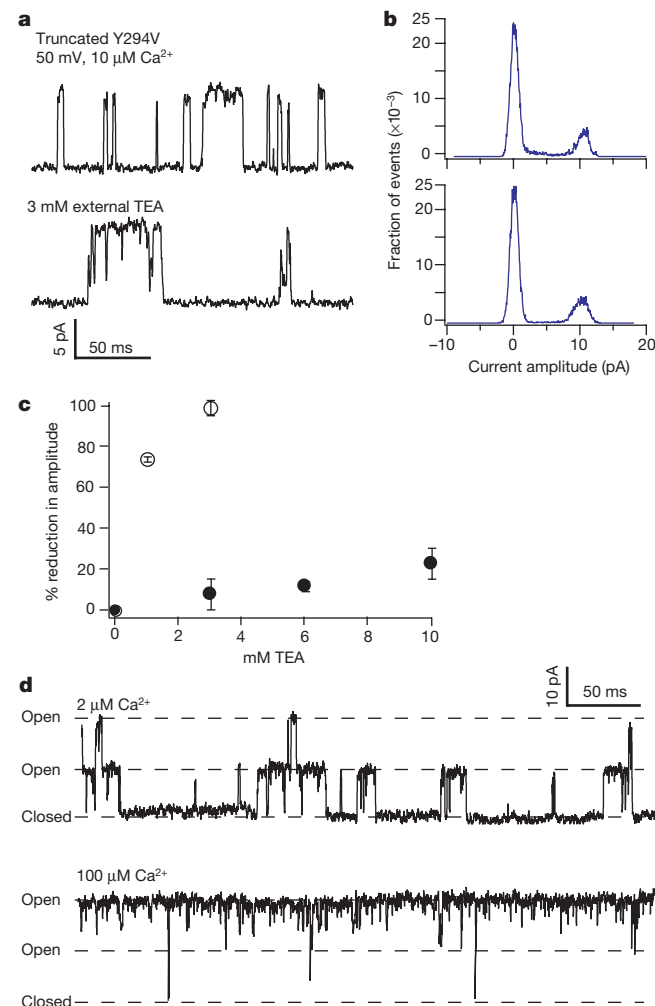


Figure 4 Data are not contaminated by endogenous BK_{Ca} channel subunits. **a**, The single-channel current from an outside-out patch before and after perfusion of 3 mM external TEA. **b**, Amplitude histograms for the stretch of data shown in **a**. **c**, Reduction in single-channel amplitude for truncated (open symbols) and truncated Tyr294Val (filled symbols) BK_{Ca} channels. **d**, Currents from an inside-out patch from an HEK 293 cell heterologously expressing truncated BK_{Ca} channels. The patch is held at 30 mV and contains two channels.

Received 6 June; accepted 12 September 2002; doi:10.1038/nature01199.

- Nelson, M. T. *et al.* Relaxation of arterial smooth muscle by calcium sparks. *Science* **270**, 633–637 (1995).
- Lewis, R. S. & Hudspeth, A. J. Voltage- and ion-dependent conductances in solitary vertebrate hair cells. *Nature* **304**, 538–541 (1983).
- Robitaille, R., Garcia, M. L., Kaczorowski, G. J. & Charlton, M. P. Functional colocalization of calcium and calcium-gated potassium channels in control of transmitter release. *Neuron* **11**, 645–655 (1993).
- Crawford, A. C. & Fettiplace, R. An electrical tuning mechanism in turtle cochlear hair cells. *J. Physiol. (Lond.)* **312**, 377–412 (1981).
- Cui, J., Cox, D. H. & Aldrich, R. W. Intrinsic voltage dependence and Ca^{2+} regulation of mslo large conductance Ca^{2+} -activated K^{+} channels. *J. Gen. Physiol.* **109**, 647–673 (1997).
- Stefani, E. *et al.* Voltage-controlled gating in a large conductance Ca^{2+} -sensitive K^{+} channel (hslo). *Proc. Natl Acad. Sci. USA* **94**, 5427–5431 (1997).
- Horrigan, F. T., Cui, J. & Aldrich, R. W. Allosteric voltage gating of potassium channels I. Mslo ionic currents in the absence of Ca^{2+} . *J. Gen. Physiol.* **114**, 277–304 (1999).
- Talukder, G. & Aldrich, R. W. Complex voltage-dependent behavior of single unliganded calcium-sensitive potassium channels. *Biophys. J.* **78**, 761–772 (2000).
- Nimigean, C. M. & Magleby, K. L. The β subunit increases the Ca^{2+} sensitivity of large conductance Ca^{2+} -activated potassium channels by retaining the gating in the bursting states. *J. Gen. Physiol.* **113**, 425–440 (1999).
- Marty, A. Ca^{2+} -dependent K^{+} channels with large unitary conductance in chromaffin cell membranes. *Nature* **291**, 497–500 (1981).
- Pallotta, B. S., Magleby, K. L. & Barrett, J. N. Single channel recordings of Ca^{2+} -activated K^{+} currents in rat muscle cell culture. *Nature* **293**, 471–474 (1981).
- Cox, D. H., Cui, J. & Aldrich, R. W. Allosteric gating of a large conductance Ca^{2+} -activated K^{+} channel. *J. Gen. Physiol.* **110**, 257–281 (1997).
- Rothberg, B. S. & Magleby, K. L. Gating kinetics of single large-conductance Ca^{2+} -activated K^{+} channels in high Ca^{2+} suggest a two-tiered allosteric gating mechanism. *J. Gen. Physiol.* **114**, 93–124 (1999).
- Rothberg, B. S. & Magleby, K. L. Voltage and Ca^{2+} activation of single large-conductance Ca^{2+} -activated K^{+} channels described by a two-tiered allosteric gating mechanism. *J. Gen. Physiol.* **116**, 75–99 (2000).
- Schreiber, M. & Salkoff, L. A novel calcium-sensing domain in the BK channel. *Biophys. J.* **73**, 1355–1363 (1997).
- Schreiber, M., Yuan, A. & Salkoff, L. Transplantable sites confer calcium sensitivity to BK channels. *Nature Neurosci.* **2**, 416–421 (1999).
- Wei, A., Solaro, C., Lingle, C. & Salkoff, L. Calcium sensitivity of BK-type K_{Ca} channels determined by a separable domain. *Neuron* **13**, 671–681 (1994).
- Bian, S., Favre, I. & Moczydlowski, E. Ca^{2+} -binding activity of a COOH-terminal fragment of the *Drosophila* BK channel involved in Ca^{2+} -dependent activation. *Proc. Natl Acad. Sci. USA* **98**, 4776–4781 (2001).
- Braun, A. F. & Sy, L. Contribution of potential EF hand motifs to the calcium-dependent gating of a mouse brain large conductance, calcium-sensitive K^{+} channel. *J. Physiol. (Lond.)* **533**, 681–695 (2001).
- Bao, L., Rapin, A. M., Holmstrand, E. C. & Cox, D. H. Elimination of the BK $_{\text{Ca}}$ channel's high-affinity Ca^{2+} sensitivity. *J. Gen. Physiol.* **120**, 173–189 (2002).
- Xia, X. M., Zeng, X. & Lingle, C. J. Multiple regulatory sites in large-conductance calcium-activated potassium channels. *Nature* **418**, 880–884 (2002).
- Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
- Bravo-Zehnder, M. *et al.* Apical sorting of a voltage- and Ca^{2+} -activated K^{+} channel α -subunit in Madin–Darby canine kidney cells is independent of N-glycosylation. *Proc. Natl Acad. Sci. USA* **97**, 13114–13119 (2000).
- Quirk, J. C. & Reinhart, P. H. Identification of a novel tetramerization domain in large conductance K_{Ca} channels. *Neuron* **32**, 13–23 (2001).
- Krause, J. D., Foster, C. D. & Reinhart, P. H. *Xenopus laevis* oocytes contain endogenous large conductance Ca^{2+} -activated K^{+} channels. *Neuropharmacology* **35**, 1017–1022 (1996).
- Shen, K. Z. *et al.* Tetraethylammonium block of Slopoke calcium-activated potassium channels expressed in *Xenopus* oocytes: evidence for tetrameric channel formation. *Pflügers Arch.* **426**, 440–445 (1994).
- Niu, X. & Magleby, K. L. Stepwise contribution of each subunit to the cooperative activation of BK channels by Ca^{2+} . *Proc. Natl Acad. Sci. USA* **99**, 11441–11446 (2002).
- Zhu, G., Zhang, Y., Xu, H. & Jiang, C. Identification of endogenous outward currents in the human embryonic kidney (HEK 293) cell line. *J. Neurosci. Methods* **81**, 73–83 (1998).
- Colquhoun, D. Binding, gating, affinity and efficacy: the interpretation of structure–activity relationships for agonists and of the effects of mutating receptors. *Br. J. Pharmacol.* **125**, 924–947 (1998).
- Horrigan, F. T. & Aldrich, R. W. Coupling between voltage-sensor activation, Ca^{2+} binding and channel opening in large conductance (BK) potassium channels. *J. Gen. Physiol.* **120**, 267–305 (2002).

Acknowledgements We thank S. Wiler for help with mutagenesis; and R. Lewis, M. Maduke, J. Trimmer, T. Middendorf, J. Sack and S. Pyott for critically reading the manuscript. R.P. was supported by a training grant from the National Institutes of Health. R.W.A. is an investigator for the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.W.A. (e-mail: raldrich@Stanford.edu).

CD4⁺CD25⁺ regulatory T cells control *Leishmania major* persistence and immunity

Yasmine Belkaid^{†‡}, Ciriaco A. Piccirillo^{‡§}, Susana Mendez^{*}, Ethan M. Shevach[§] & David L. Sacks^{*}

^{*} Laboratory of Parasitic Diseases and [§] Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

[‡] These authors contributed equally to this work

The long-term persistence of pathogens in a host that is also able to maintain strong resistance to reinfection, referred to as concomitant immunity, is a hallmark of certain infectious diseases, including tuberculosis and leishmaniasis. The ability of pathogens to establish latency in immune individuals often has severe consequences for disease reactivation^{1–3}. Here we show that the persistence of *Leishmania major* in the skin after healing in resistant C57BL/6 mice is controlled by an endogenous population of CD4⁺CD25⁺ regulatory T cells. These cells constitute 5–10% of peripheral CD4⁺ T cells in naive mice and humans, and suppress several potentially pathogenic responses *in vivo*, particularly T-cell responses directed against self-antigens⁴. During infection by *L. major*, CD4⁺CD25⁺ T cells accumulate in the dermis, where they suppress—by both interleukin-10-dependent and interleukin-10-independent mechanisms—the ability of CD4⁺CD25[−] effector T cells to eliminate the parasite from the site. The sterilizing immunity achieved in mice with impaired IL-10 activity is followed by the loss of immunity to reinfection, indicating that the equilibrium established between effector and regulatory T cells in sites of chronic infection might reflect both parasite and host survival strategies.

After inoculation of 1,000 metacyclic promastigotes of *L. major* into the ear dermis, C57BL/6 mice develop a small lesion that resolves spontaneously after 8–10 weeks. Self-cure is associated with the killing of more than 99% of the parasites in the skin, and the animals develop long-lasting immunity to secondary challenge. Despite this high level of acquired resistance, a few viable organisms (10^2 – 10^4) persist in the site of the former skin lesion and in the draining lymph node (LN) (data not shown). The chronically infected dermis is characterized by the maintenance of a large number of lymphocytes, mainly CD4⁺ T cells ($2.2 \times 10^5 \pm 1.8 \times 10^4$ per ear compared with $9.3 \times 10^3 \pm 2.3 \times 10^3$ in a naive ear under steady-state conditions; data not shown). Of the CD4⁺ T cells in the chronic site, 40–50% expressed CD25 and CTLA-4 and were uniformly low for CD45RB (Fig. 1a). In contrast, the dermal CD4⁺CD25[−] T cells were low or negative for CTLA-4, and 20% of them were CD45RB^{high}. Freshly explanted CD4⁺CD25⁺ T cells from chronically infected dermis were non-responsive when stimulated *in vitro* with soluble anti-CD3 but responded to stimulation of the T-cell antigen receptor (TCR) in the presence of interleukin (IL)-2 (Fig. 1b). The cells potentially suppressed anti-CD3-induced proliferation of lesion-derived CD4⁺CD25[−] T cells (Fig. 1c) that was not reversed by the addition of anti-IL-10 receptor (anti-IL-10R) or anti-TGF- β (see Supplementary Information). These findings demonstrate that the CD4⁺CD25⁺ T cells accumulating within chronic sites of *L. major* infection in the skin are phenotypically and functionally identical to the naturally occurring CD4⁺CD25⁺ regulatory T cells^{5–7}. Because the lesion-derived CD4⁺CD25[−] T cells exhibited low levels of antigen-specific proliferation, suppression of

[†] Present address: Research Foundation, Division of Molecular Immunology, Children's Hospital, 3333 Burnet Avenue, Cincinnati, Ohio, MLC 7021 45229, USA.

antigen-specific responses was examined in the context of effector cytokine production. When lesion-derived $CD4^+CD25^-$ T cells were incubated with $CD4^+CD25^+$ T cells in the presence of *L. major*-infected macrophages, significant (50%) suppression of interferon (IFN)- γ production was observed that was not reversed by the addition of anti-IL-10R (Fig. 1d).

Although IL-10 did not seem to have a role in lesion-derived $CD4^+CD25^+$ -mediated suppression of T-cell proliferation or in the inhibition of antigen-specific IFN- γ production by $CD4^+CD25^-$ T cells *in vitro*, $CD4^+CD25^+$ T cells from naive mice do express elevated levels of IL-10 mRNA and can secrete IL-10 when stimulated *in vitro*^{6,8}. Because IL-10 has a crucial role in the failure of the infected host to clear *L. major* infection⁹, it was of interest to determine whether the lesion-derived $CD4^+CD25^+$ T cells could produce IL-10 in response to *L. major*-infected fetal-skin-derived dendritic cells (DCs), conditions we have previously found to optimize antigen-specific cytokine secretion by unfractionated $CD4^+$ T cells *in vitro*⁹. Whereas $CD4^+CD25^-$ T cells released large amounts of IFN- γ ($14,000 \pm 520$ pg ml⁻¹) but no IL-10, the $CD4^+CD25^+$ T cells released a large amount of IL-10 ($1,070 \pm 5.8$ pg ml⁻¹) and a small amount of IFN- γ (800 ± 6.3 pg ml⁻¹) (Fig. 1e). No IL-5 and only a marginal amount of IL-4 (<45 pg ml⁻¹) were detected, indicating that $CD4^+CD25^+$ T cells from the chronic site are functionally distinct from T helper type 2 cells. The uninfected DCs failed to activate either subset to

produce cytokine, and no cytokine was detected when naive dermal lymphocytes were incubated with infected DCs (data not shown). Finally, the *L. major*-induced IL-10 production by lesion-derived $CD4^+CD25^+$ T cells was not a consequence of non-specific activation of DCs, involving increased presentation of self-antigens and/or expression of co-stimulatory molecules, because in a separate experiment no IL-10 was detectable when DCs were activated with agonist anti-CD40 or with DCs infected with *Toxoplasma gondii*, whereas in response to the *L. major*-infected DCs, the $CD4^+CD25^+$ T cells produced 506 ± 133 pg ml⁻¹ IL-10 (data not shown).

The capacity of lesion-derived $CD4^+CD25^+$ and $CD4^+CD25^-$ T-cell subsets to produce IL-10 and IFN- γ , respectively, raised the possibility that a delicate balance exists between suppressor and effector T-cell functions within the chronically infected skin. To analyse the function of these two subsets further, they were adoptively transferred separately or together into RAG^{-/-} mice 1 day before intradermal (i.d.) challenge with *L. major*. At 3.5 weeks, just before the normal onset of parasite clearance¹⁰, infected wild-type (WT) mice harboured a high parasitic load in the dermis (Fig. 2a) and draining LN (data not shown), comparable to that in infected RAG^{-/-} mice. The dermis of WT mice contained 2×10^5 $CD4^+$ T cells, 22% of which expressed CD25 and 4.7% of which released IFN- γ after exposure to infected DCs (Fig. 2c, and see Supplementary Information). RAG^{-/-} recipient mice of lesion-derived $CD4^+CD25^-$ T cells were highly resistant to infection and rapidly

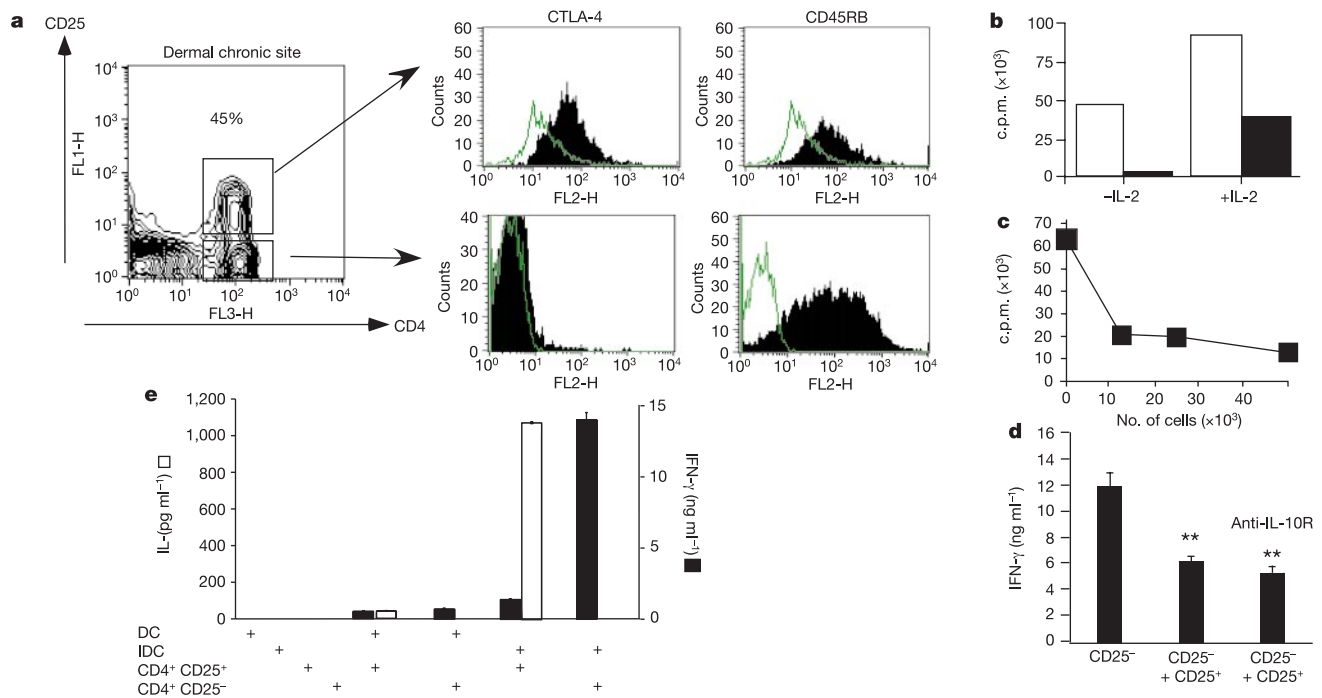


Figure 1 $CD4^+CD25^+$ T cells from chronic lesions are suppressive and secrete IL-10 in response to *L. major* *in vitro*. **a**, C57BL/6 mice (Division of Cancer Treatment, National Cancer Institute) were infected intradermally with 1,000 *L. major* metacyclic promastigotes (V1, MHOM/IL/80/Friedlin) as described previously¹⁰. Dermal cells from chronically infected C57BL/6 ears (22 weeks) were collected, stained and analysed by flow cytometry as described previously⁹. Histograms are gated on TCR β , $CD4^+$, $CD25^+$ or $CD25^-$ cells; CTLA-4 staining was performed after cell permeabilization; the isotype control is shown as a light line. **b**, $CD4^+CD25^+$ (black bars) and $CD4^+CD25^-$ (white bars) T cells from chronically infected sites were separated by cell sorting as previously described⁶. Proliferation assays were performed as described previously⁶ by stimulating $CD4^+CD25^-$ or $CD4^+CD25^+$ T cells (5×10^4) from chronic lesions with anti-CD3 and T-depleted irradiated spleen cells (APCs, $(1-2) \times 10^5$), with or without exogenous recombinant human IL-2. **c**, Proliferation assays were performed by stimulating $CD4^+CD25^-$ T cells (5×10^4) from chronic lesions with anti-CD3 (0.5μ g/ml) and APC

($1-2 \times 10^5$) in the absence or presence of increasing number of $CD4^+CD25^+$ T cells from chronic lesions. **d**, Peritoneal macrophages (2.5×10^4) were infected with serum-opsonized *L. major* metacyclic promastigotes (two parasites per cell) and used to stimulate $CD4^+CD25^-$ T cells either alone (2.5×10^4) or with lesion-derived $CD4^+CD25^+$ (2.5×10^4) T cells. The co-culture was performed in the presence or absence of anti IL-10R (clone 1B1.3a; 50μ g/ml DNAX²⁴). IFN- γ production was analysed by ELISA (Endogen). Values are means $\pm$ s.d. of triplicates per condition. Cytokine production was significantly decreased ($**P < 0.001$) compared with CD25 alone. **e**, $CD4^+CD25^+$ and $CD4^+CD25^-$ T cells from chronically infected dermis (20 weeks after challenge) were incubated for 48 h with uninfected (DC) or *L. major*-infected (IDC) fetal-skin-derived dendritic cells²⁵ at a ratio of 5 lymphocytes per APC. Cytokine production was analysed by ELISA (Endogen). Values are means $\pm$ s.d. of triplicates per condition. The results are representative of four experiments.

controlled parasite growth (10^4 -fold decrease compared with un-reconstituted RAG $^{-/-}$ mice) (Fig. 2a). This resistance was associated with a large number of intradermal lymphocytes (threefold the total number initially transferred) (Fig. 2c), 25% of which produced antigen-specific IFN- γ , and only 5% of which expressed CD25 (see Supplementary Information). Mice transferred with CD4 $^{+}$ CD25 $^{+}$ T cells alone were highly susceptible to infection, both in terms of parasite growth and lesion size (Fig. 2a and b). Their non-healing phenotype was associated at 3.5 weeks with a small number of CD4 $^{+}$ lymphocytes found in the infected skin (1.2×10^4 cells per ear) (Fig. 2c), 46% of which still expressed CD25. Transfer of CD4 $^{+}$ CD25 $^{+}$ together with CD4 $^{+}$ CD25 $^{-}$ T cells (at the 1:10 ratio found in naive mice) suppressed (1) the ability of the CD4 $^{+}$ CD25 $^{-}$ cells to mediate early control of parasite growth in the site (Fig. 2a), (2) the rapid accumulation of lymphocytes in the skin (Fig. 2c), and (3) the frequency of IFN- γ producing dermal CD4 $^{+}$ T cells (Fig. 2c and Supplementary Information).

Infected RAG $^{-/-}$ mice developed a delayed and minor pathology despite their high parasite burden, owing to the absence of T cells capable of mediating the acute inflammatory response to infection. WT C57BL/6 mice developed a small lesion that began to resolve spontaneously after 7 weeks. RAG $^{-/-}$ mice reconstituted only with CD4 $^{+}$ CD25 $^{-}$ T cells developed a rapid and transient inflammatory response that coincided with the rapid killing of the parasites in the site, and by 6 weeks they showed no overt pathology (Fig. 2b). RAG $^{-/-}$ recipients of CD4 $^{+}$ CD25 $^{+}$ T cells alone developed a non-healing lesion that evolved more rapidly than the lesion in the RAG $^{-/-}$ controls. Thus, CD4 $^{+}$ CD25 $^{+}$ T cells markedly exacerbated lesion development and parasite growth even though the cells were obtained from mice that were themselves highly resistant to challenge infection. Mice to which both CD4 $^{+}$ CD25 $^{+}$ and CD4 $^{+}$ CD25 $^{-}$ T cells had been transferred developed a minor pathology, suggesting an immune phenotype similar to that of WT mice.

Because the leukopenic environment in RAG $^{-/-}$ mice has been

shown to favour a rapid expansion of transferred lymphocytes, we also confirmed that WT C57BL/6 mice transferred with CD4 $^{+}$ CD25 $^{+}$ T cells purified from a chronic site became highly susceptible to *L. major* infection (see Supplementary Information). The non-healing phenotype transferred into intact animals by the CD4 $^{+}$ CD25 $^{+}$ T cells was not associated with a T helper type 2 (IL-4 or IL-5) response induced by these cells (data not shown).

As an initial approach to determining the origin of the lesion-derived CD4 $^{+}$ CD25 $^{+}$ regulatory T cells, CD4 $^{+}$ CD25 $^{+}$ and CD4 $^{+}$ CD25 $^{-}$ T cells purified from lymphoid organs of naive C57BL/6 mice were transferred separately or together into RAG $^{-/-}$ recipients 1 day before i.d. challenge with *L. major*. RAG $^{-/-}$ recipients of naive CD4 $^{+}$ CD25 $^{-}$ T cells were highly resistant to infection; at 10 weeks after infection, they had a 10^4 – 10^5 reduction compared with RAG $^{-/-}$ controls in the number of parasites in the site, and surprisingly developed little or no overt pathology (Fig. 2d, e). This strong resistance was associated with a large number of IFN- γ -producing CD4 $^{+}$ T cells in the site (3.2×10^4) (Fig. 2f). A small percentage (6%) of the dermal lymphocytes derived from the CD4 $^{+}$ CD25 $^{-}$ T cells became CD25 $^{+}$ (see Supplementary Information). RAG $^{-/-}$ recipients of CD4 $^{+}$ CD25 $^{+}$ T cells were highly susceptible to *L. major* infection; at 10 weeks they harboured massive numbers of parasites in the site and developed a non-healing pathology that was more severe than in non-reconstituted RAG $^{-/-}$ mice. A small number of lymphocytes (1.3×10^4 CD4 $^{+}$ TCR- β^{+} cells per ear) were present in the dermis, 56% of which were still CD25 $^{+}$ (see Supplementary Information), but only 1.4×10^2 of which were able to produce IFN- γ in response to *Leishmania* antigen *in vitro*. RAG $^{-/-}$ recipients of both naive populations, again at a 1:10 ratio of CD25 $^{+}$:CD25 $^{-}$, controlled parasite growth but contained a significantly higher parasite number 10 weeks after infection than RAG $^{-/-}$ -receiving CD4 $^{+}$ CD25 $^{-}$ T cells alone, and had 4-fold fewer IFN- γ -producing CD4 $^{+}$ cells in the dermis (7.4×10^3). Finally, they developed small healing lesions comparable to those seen in the WT mice.

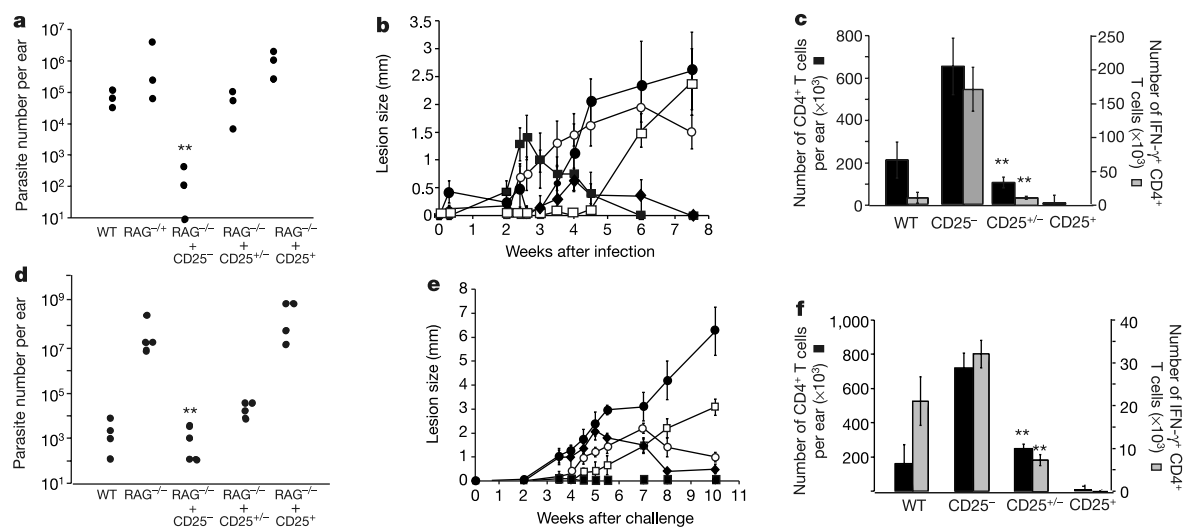


Figure 2 Lesion-derived CD4 $^{+}$ CD25 $^{+}$ T cells suppress *L. major* immunity mediated by CD4 $^{+}$ CD25 $^{-}$ T cells *in vivo*. CD4 $^{+}$ CD25 $^{-}$ or CD4 $^{+}$ CD25 $^{+}$ T cells were purified from the dermis of infected mice (22 weeks) (a–c) or the lymph node of naive mice (d–f); 2×10^5 CD4 $^{+}$ CD25 $^{+}$ or 2×10^5 CD4 $^{+}$ CD25 $^{-}$ T cells alone or in combination (2×10^5 CD4 $^{+}$ CD25 $^{-}$ and 2×10^4 CD4 $^{+}$ CD25 $^{+}$) (a–c) and 4×10^5 CD4 $^{+}$ CD25 $^{+}$ or 4×10^5 CD4 $^{+}$ CD25 $^{-}$ T cells alone or in combination (4×10^5 CD4 $^{+}$ CD25 $^{-}$ and 4×10^4 CD4 $^{+}$ CD25 $^{+}$) (d–f) were transferred intravenously into C57BL/6 RAG $^{-/-}$ mice 1 d before i.d. challenge with 1,000 *L. major* metacysts. **a, d**, Number of parasites per ear at 3.5 (a) or 10 (d) weeks after challenge. The parasite number in the CD4 $^{+}$ CD25 $^{-}$ transferred group is significantly less (asterisks, $P < 0.001$) than in all other groups. Three mice per

group were analysed. **b, e**, Diameters of indurations in WT (open circles) or RAG $^{-/-}$ (open squares) recipients of CD4 $^{+}$ CD25 $^{+}$ cells (filled circles), CD4 $^{+}$ CD25 $^{-}$ cells (filled squares) or CD4 $^{+}$ CD25 $^{+}$ plus CD4 $^{+}$ CD25 $^{-}$ T cells (filled diamonds). Values are means $\pm$ s.d. of induction, 4–10 ears per time point. **c, f**, Absolute numbers of TCR- β^{+} CD4 $^{+}$ T cells and IFN- γ^{+} TCR- β^{+} CD4 $^{+}$ T cells recovered from the dermis of the different groups of mice 3.5 (c) and 10 (f) weeks after transfer and infection. For intracellular staining, dermal cells were incubated with bone-marrow-derived dendritic cells²⁶ infected as described previously⁹ before being labelled for surface markers and IFN- γ . Ears were treated individually and values shown are numbers of positive cells per ear (means $\pm$ s.d.), six ears per point. **, $P < 0.001$.

Although the above studies clearly demonstrate that $CD4^+CD25^+$ T cells derived from normal mice can suppress *L. major* immunity, they do not directly address the origin of the suppressor T cells found in the infected dermis, as the CD25 marker might be gained or lost after the transfer of $CD4^+CD25^-$ and $CD4^+CD25^+$ T cells to $RAG^{-/-}$ recipients. We therefore transferred, at a 1:10 ratio, $CD4^+CD25^+$ T cells from naive $Ly5.1^+$ mice together with $CD4^+CD25^-$ T cells from naive $Ly5.2^+$ mice into $RAG^{-/-}$ recipients. By 3 weeks after infection, during the accumulation of the peak number of parasites in the site (Fig. 3d), 48% of the $CD4^+$ T cells present in the dermis expressed $Ly5.1$ (Fig. 3a), but only 13% of the $Ly5.1^+$ cells still expressed CD25 (Fig. 3b). The early and preferential recruitment of regulatory T cells to the site of infection might be crucial to the delayed onset of immunity that occurs during *L. major* infection¹⁰. At later time points, during the stage of lesion formation and immune clearance of parasites from the site (Fig. 3d), the ratio of dermal $TCR-\beta^+CD4^+$ T cells that were $Ly5.1^+$ decreased to 27% at 4 weeks, and to 12% at 7 weeks. This change was due almost entirely to the expansion and recruitment of $Ly5.2^+CD4^+$ T cells in the site and not to a decrease in the absolute number of $Ly5.1^+CD4^+$ T cells (Fig. 3a). At 9 weeks, at about the onset of the chronic phase, the $Ly5.1^+$ cells again represented more than half of the $TCR-\beta^+CD4^+$ T cells in the site (63%). When the cells present in the chronic lesion at 9 weeks were stimulated with

infected DCs, the vast majority (90%) of the cells able to release $IFN-\gamma$ in response to *L. major* were found in the $Ly5.2^+$ population (Fig. 3c). Throughout the course of infection, no more than 4% of the $Ly5.2^+$ cells expressed CD25 (Fig. 3b). Thus, few of the transferred $CD25^-$ T cells became $CD25^+$ and the vast majority of the $CD4^+CD25^+$ T cells that homed initially to the inoculation site, and were maintained during the chronic phase of the infection, originated from an endogenous $CD4^+CD25^+$ regulatory T-cell population. Taken together, the cell transfer studies with naive T-cell subsets establish that from within the compartment of $CD4^+CD25^+$ T cells present in the naive animal, a population of suppressor cells is selectively recruited to the site of *L. major* challenge in the skin, where they promote the establishment of infection and ensure the long-term survival of the parasite in the immune host.

To establish the relative contribution of IL-10 produced by regulatory T cells in parasite persistence, and to address other potential suppressor mechanisms involved in the evolution of the chronic infection, naive $CD4^+CD25^-$ T cells were transferred alone or in combination with either $CD4^+CD25^+$ T cells from naive WT (C57BL/10) or from $IL-10^{-/-}$ C57BL/10 mice ($CD4^+CD25^{+10^0}$) into $RAG^{-/-}$ recipients. At 3.5 weeks after infection, the approximate time of peak parasite load establishment in C57BL/10, recipients of $CD4^+CD25^-$ T cells alone harboured only 1.2×10^3

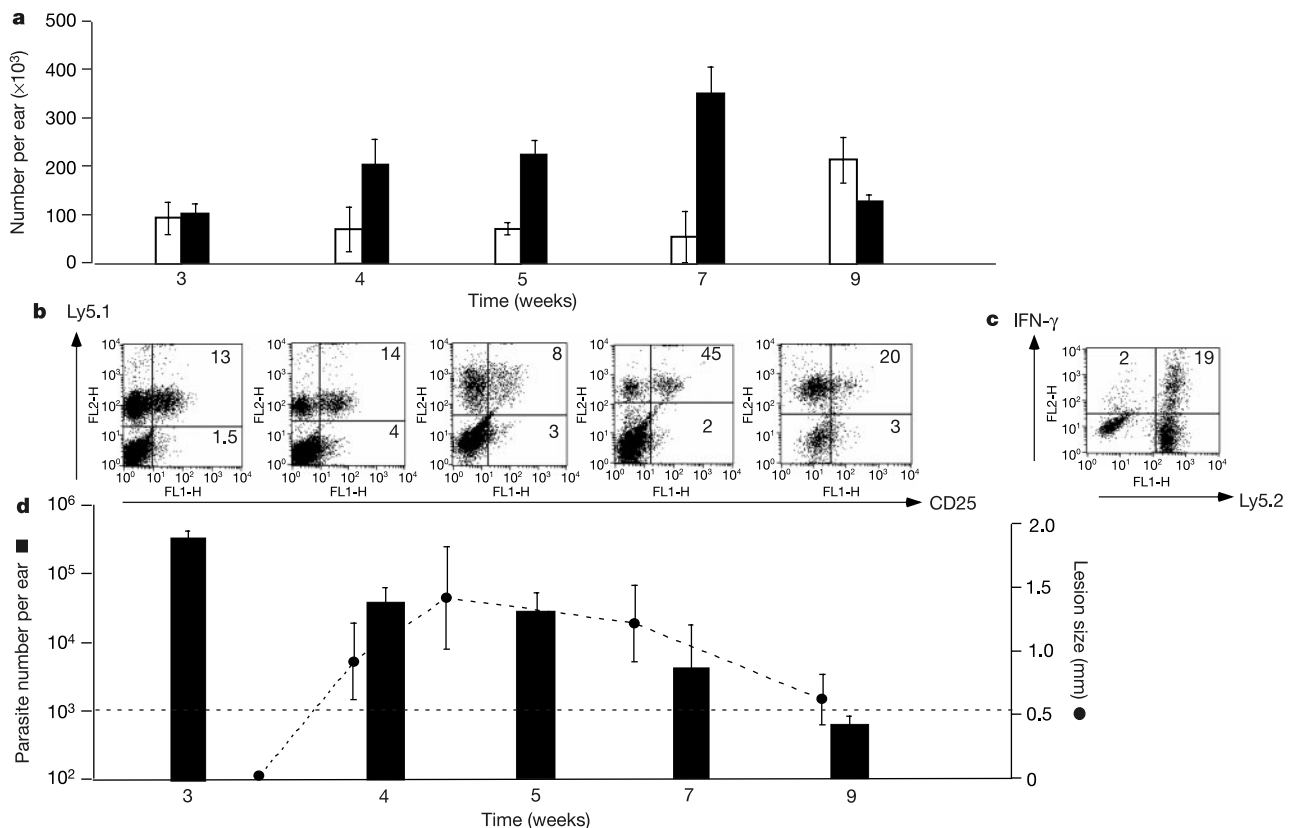


Figure 3 $CD4^+CD25^+$ T cells accumulating in chronic sites originate from naturally occurring $CD4^+CD25^+$ regulatory T cells. Naive lymph node $CD4^+CD25^+$ and $CD4^+CD25^-$ T cells from $Ly5.1$ congenic (B6.SJL CD45a, $Ly5a$, Taconic) and $Ly5.2$ WT C57BL/6 mice, respectively, were purified and co-injected intravenously at a 1:10 ratio of $CD4^+CD25^+$ to $CD4^+CD25^-$ (3×10^4 and 3×10^5 , respectively) into $RAG^{-/-}$ mice, 1 d before i.d. infection. **a**, Absolute numbers of $TCR-\beta^+CD4^+Ly5.1^+$ (open bars) and $Ly5.2^+$ (filled bars) cells per ear during the course of the infection. Values are means $\pm$ s.d. of four to six ears per time point. **b**, Expression of $Ly5.1/CD25$ markers on $TCR-\beta$, $CD4^+$ gated events. Dermal cells from individual ears were extracted and stained with anti- $TCR-\beta$ chain, $CD4$, $Ly5.1$ and $CD25$. The number in each quadrant represents

the percentage of positive events in the quadrant. This panel is representative of four separated analyses. **c**, $IFN-\gamma$ production by dermal cells (9 weeks after transfer/challenge) after exposure *in vitro* to *L. major*-infected bone-marrow-derived dendritic cells. Cells were gated on $CD4^+TCR-\beta^+$ events. Values refer to the percentages of $CD4^+TCR-\beta^+$ lymphocytes that are either $Ly5.2^+/IFN-\gamma$ or $Ly5.2^-/IFN-\gamma$ compared with the isotype control. This panel is representative of four separate analyses. **d**, Parasite number and lesion size during the course of infection in transferred mice. Solid bars are means $\pm$ s.d. of parasites per ear, four ears per time point. Filled circles are means $\pm$ s.d. of the induration (4–20 ears per time point). The results are representative of three separate experiments.

parasites in the site (Fig. 4a), and as observed previously (Fig. 2e) they developed a transient and minimal dermal pathology (data not shown). Recipients transferred together with either $CD4^+CD25^-$ / $CD4^+CD25^+$ or $CD4^+CD25^-/CD4^+CD25^{+10^0}$ contained comparable numbers of parasites in the skin (3.1×10^4 versus 4.1×10^4) at 3.5 weeks, which in each case was significantly greater than the dermal parasite loads in recipients of $CD4^+CD25^-$ alone ($P < 0.001$). At this stage of infection, the $CD4^+CD25^+$ T cells might modulate effector T-cell¹¹ or antigen-presenting cell (APC) functions¹² by means of the cell-contact-dependent mechanisms that are seen *in vitro*, or through the secretion of suppressive cytokines other than IL-10, such as TGF- β ¹³. At 17 weeks, when chronic stage infections were well established, recipients of $CD4^+CD25^-$ T cells alone had completely eliminated parasites from the site (Fig. 4a). Recipients of $CD4^+CD25^-/CD4^+CD25^+$ T cells from WT mice developed a chronic lesion similar to that seen in WT mice with more than 1,000 parasites persisting in the site. In contrast, recipients of $CD4^+CD25^-/CD4^+CD25^{+10^0}$ T cells healed faster and completely cleared the parasites from the site. This result confirms that IL-10 produced by regulatory T cells contributes directly to parasite persistence by either modulating APC function, inhibiting cytokine production by T helper type 1 cells or rendering macrophages refractory to activation by IFN- γ for intracellular killing¹⁴. Although IL-10 produced by $CD4^+CD25^+$ regulatory T cells is essential to the establishment of persistent infection, early in the infectious process $CD4^+CD25^+$ T cells can promote parasite survival and growth in an IL-10-independent manner.

We were interested to know whether the durable immunity to reinfection that is a hallmark of acquired resistance in healed mice and humans is dependent on parasite persistence. Because RAG^{-/-} recipients of $CD25^+$ -depleted $CD4^+$ T cells developed autoimmune pathologies and failed to thrive, the re-challenge studies were performed in IL-10^{-/-} C57BL/10 mice, and in C57BL/6 WT mice that had been treated during the chronic stage of their primary infection with anti-IL-10R antibody, conditions that also result in the complete clearance of parasites from the skin and draining LN⁹. Reinfection of IL-10^{-/-} mice in the contralateral, uninjected ear resulted in parasitic loads at 5 weeks that were slightly greater than those present in the primary challenge site of the naive IL-10^{-/-}

mice (Fig. 4b). In contrast, re-challenge of the chronically infected C57BL/10 WT mice resulted in a 10^2 – 10^3 -fold reduction in dermal parasite loads compared with naive WT mice. Similarly, reinfection of anti-IL-10R-treated C57BL/6 mice in the uninjected ear resulted in parasitic loads at 5 weeks that were comparable to those established after primary infection in naive mice, whereas healed mice treated with control antibody maintained powerful immunity to reinfection. The control treated, healed mice were also completely protected against the development of dermal lesions, whereas the anti-IL-10R-treated mice developed lesions after reinfection that were as severe as those after primary challenge in the naive mice (data not shown). Thus, the maintenance of a residual source of infection secondary to IL-10 production by $CD4^+CD25^+$ T cells at the lesion site is required for the preservation of acquired immunity to *L. major*.

The present studies establish that the cells responsible for *L. major* persistence after resolution of dermal lesions in C57BL/6 mice are phenotypically and functionally indistinguishable from the endogenous $CD4^+CD25^+$ regulatory T cells that have been described in the context of peripheral tolerance and autoimmunity⁵. During the chronic phase of *L. major* infection, the small number of parasites persisting in the dermis can be efficiently transmitted back to their vector sandflies^{9,15}. Thus, the expansion and/or recruitment of regulatory T cells to the site of *L. major* infection might reflect an adaptive strategy by the parasite to maintain its transmission cycle in nature. These data raise the possibility that the more severe, non-healing forms of leishmaniasis, which in humans are associated with the overproduction of IL-10 (refs 16, 17), might be due to an imbalance in the number and activity of parasite-driven regulatory T cells. In the prototypic BALB/c model of non-healing cutaneous leishmaniasis, the $CD4^+$ cells that suppressed *L. major* immunity were found within an IL-4/IL-10-producing $CD45RB^{low}$ population that also inhibited colitis¹⁸.

A high proportion of $CD4^+$ T cells able to release IL-10 can be found in other chronic infections, such as tuberculosis¹⁹, malaria²⁰, toxoplasmosis²¹ and HIV²², again indicating the presence of regulatory T cells. The deviation of a principal immune regulatory mechanism in the host might therefore be a general strategy for pathogens to favour their long-term survival. Alternatively, *L. major*

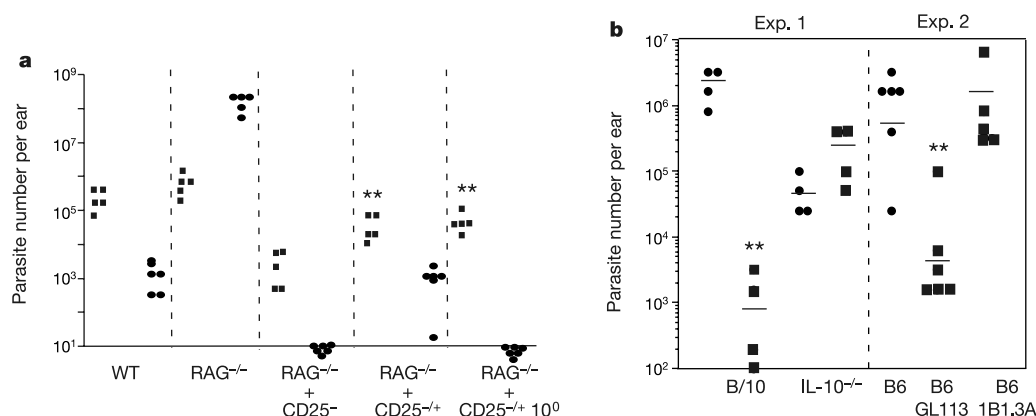


Figure 4 IL-10 produced by $CD4^+CD25^+$ T cells is required for parasite persistence and for the maintenance of immunity to reinfection. **a**, WT (C57BL/10, Taconic) $CD4^+CD25^-$ T cells (4×10^5) were transferred into RAG^{-/-} (Taconic) mice either alone or in combination with $CD4^+CD25^+$ T cells (4×10^4) from WT or IL-10^{-/-} (10^6) mice (Taconic). Numbers of parasites in the dermis of C57BL/10, RAG^{-/-} or reconstituted RAG^{-/-}, 3.5 weeks (squares) or 17 weeks (circles) after transfer/infection. Parasite number is significantly greater (asterisks, $P < 0.001$) than in $CD4^+CD25^-$ recipient mice at the same time point; five or six ears per point. The results are representative of two separate experiments. **b**, In experiment 1, C57BL/10 WT or IL-10^{-/-} mice were infected in one ear with 1,000 *L. major* promastigotes. Six months after infection, healed (squares)

and naive control (circles) mice were challenged in the other ear with 1,000 *L. major* promastigotes. Values are parasite numbers in individual ears 5 weeks after challenge. Results are representative of two separate experiments. In experiment 2, C57BL/6 mice infected in one ear for 7 months were treated with isotype control (GL113) or anti-IL-10R antibody (DNAX, 1B1.3A, 0.5 mg intraperitoneally every 3 days) for 2 weeks. Two months after treatment, healed mice (squares) or untreated naive (circles) mice were challenged in the other ear with 1,000 *L. major* promastigotes. Values are numbers of parasites in individual ears 5 weeks after challenge. Bars represent the geometric mean parasite number, four to six ears per group. **, $P < 0.001$.

might simply be exploiting a normal homeostatic process that protects the host from tissue damage associated with powerful immune responses in the skin. Although one consequence of this regulation is chronic infection and the potential for disease reactivation, parasite persistence itself provides a major benefit to the host by maintaining lifelong immunity to reinfection. Our findings in IL-10^{-/-} mice and in anti-IL-10R-treated WT mice are in agreement with previous work²³ indicating that in mice that have been manipulated to achieve sterile cure, protection against secondary challenge is lost. Thus, regulatory T cells seem to produce the paradoxical effects of suppressing effector T-cell functions locally while maintaining a recirculating pool of tissue-seeking memory cells that confer powerful immunity to reinfection. Undoubtedly, the equilibrium that is established between effector and regulatory T cells in sites of chronic infection reflects the co-evolution of host and parasite survival strategies, and raises an intriguing question of whether regulatory T cells should be avoided or co-opted by vaccination.

Received 14 May; accepted 16 September 2002; doi:10.1038/nature01152.

1. el Hassan, A. M. *et al.* Post kala-azar dermal leishmaniasis in the Sudan: clinical features, pathology and treatment. *Trans. R. Soc. Trop. Med. Hyg.* **86**, 245–248 (1992).
2. Momeni, A. Z. & Aminjavaheri, M. Clinical picture of cutaneous leishmaniasis in Isfahan, Iran. *Int. J. Dermatol.* **33**, 260–265 (1994).
3. Alvar, J. *et al.* Leishmania and human immunodeficiency virus coinfection: the first 10 years. *Clin. Microbiol. Rev.* **10**, 298–319 (1997).
4. Shevach, E. M. CD4+ CD25+ suppressor T cells: more questions than answers. *Nature Rev. Immunol.* **2**, 389–400 (2002).
5. Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J. Immunol.* **155**, 1151–1164 (1995).
6. Thornton, A. M. & Shevach, E. M. CD4+CD25+ immunoregulatory T cells suppress polyclonal T cell activation in vitro by inhibiting interleukin 2 production. *J. Exp. Med.* **188**, 287–296 (1998).
7. Piccirillo, C. A. *et al.* CD4+CD25+ regulatory T cells can mediate suppressor function in the absence of transforming growth factorβ1 production and responsiveness. *J. Exp. Med.* **196**, 237–246 (2002).
8. Papiernik, M., de Moraes, M. L., Pontoux, C., Vasseur, F. & Penit, C. Regulatory CD4 T cells: expression of IL-2R alpha chain, resistance to clonal deletion and IL-2 dependency. *Int. Immunol.* **10**, 371–378 (1998).
9. Belkaid, Y. *et al.* The role of interleukin (IL)-10 in the persistence of *Leishmania major* in the skin after healing and the therapeutic potential of anti-IL-10 receptor antibody for sterile cure. *J. Exp. Med.* **194**, 1497–1506 (2001).
10. Belkaid, Y. *et al.* A natural model of *Leishmania major* infection reveals a prolonged 'silent' phase of parasite amplification in the skin before the onset of lesion formation and immunity. *J. Immunol.* **165**, 969–977 (2000).
11. Piccirillo, C. A. & Shevach, E. M. Cutting edge: control of CD8+ T cell activation by CD4+CD25+ immunoregulatory cells. *J. Immunol.* **167**, 1137–1140 (2001).
12. Cederbom, L., Hall, H. & Ivars, F. CD4+CD25+ regulatory T cells down-regulate co-stimulatory molecules on antigen-presenting cells. *Eur. J. Immunol.* **30**, 1538–1543 (2000).
13. Read, S., Malmstrom, V. & Powrie, F. Cytotoxic T lymphocyte-associated antigen 4 plays an essential role in the function of CD25+CD4+ regulatory cells that control intestinal inflammation. *J. Exp. Med.* **192**, 295–302 (2000).
14. Gazzinelli, R. T., Oswald, I. P., James, S. L. & Sher, A. IL-10 inhibits parasite killing and nitrogen oxide production by IFN-γ-activated macrophages. *J. Immunol.* **148**, 1792–1796 (1992).
15. Mendez, S. *et al.* The potency and durability of DNA- and protein-based vaccines against *Leishmania major* evaluated using low-dose, intradermal challenge. *J. Immunol.* **166**, 5122–5128 (2001).
16. Karp, C. L. *et al.* In vivo cytokine profiles in patients with kala-azar. Marked elevation of both interleukin-10 and interferon-gamma. *J. Clin. Invest.* **91**, 1644–1648 (1993).
17. Gasim, S. *et al.* High levels of plasma IL-10 and expression of IL-10 by keratinocytes during visceral leishmaniasis predict subsequent development of post-kala-azar dermal leishmaniasis. *Clin. Exp. Immunol.* **111**, 64–69 (1998).
18. Powrie, F., Correa-Oliveira, R., Mauze, S. & Coffman, R. L. Regulatory interactions between CD45RB^{high} and CD45RB^{low} CD4+ T cells are important for the balance between protective and pathogenic cell-mediated immunity. *J. Exp. Med.* **179**, 589–600 (1994).
19. Gerosa, F. *et al.* CD4+ T cell clones producing both interferon-gamma and interleukin-10 predominate in bronchoalveolar lavages of active pulmonary tuberculosis patients. *Clin. Immunol.* **92**, 224–234 (1999).
20. Plebanski, M. *et al.* Interleukin 10-mediated immunosuppression by a variant CD4 T cell epitope of *Plasmodium falciparum*. *Immunity* **10**, 651–660 (1999).
21. Jankovic, D. *et al.* In the absence of IL-12, CD4+ T cell responses to intracellular pathogens fail to default to a Th2 pattern and are host protective in an IL-10^{-/-} setting. *Immunity* **16**, 429–439 (2002).
22. Ostrowski, M. A. *et al.* Quantitative and qualitative assessment of human immunodeficiency virus type 1 (HIV-1)-specific CD4+ T cell immunity to gag in HIV-1-infected individuals with differential disease progression: reciprocal interferon-gamma and interleukin-10 responses. *J. Infect. Dis.* **184**, 1268–1278 (2001).
23. Uzonza, J. E., Wei, G., Yurkowski, D. & Bretscher, P. Immune elimination of *Leishmania major* in mice: implications for immune memory, vaccination, and reactivation disease. *J. Immunol.* **167**, 6967–6974 (2001).
24. O'Farrell, A. M., Liu, Y., Moore, K. W. & Mui, A. L. IL-10 inhibits macrophage activation and proliferation by distinct signaling mechanisms: evidence for Stat3-dependent and -independent pathways. *EMBO J.* **17**, 1006–1018 (1998).
25. von Stebut, E., Belkaid, Y., Jakob, T., Sacks, D. L. & Udey, M. C. Uptake of *Leishmania major* amastigotes results in activation and interleukin 12 release from murine skin-derived dendritic cells: implications for the initiation of anti-*Leishmania* immunity. *J. Exp. Med.* **188**, 1547–1552 (1998).
26. Lutz, M. B. *et al.* An advanced culture method for generating large quantities of highly pure dendritic cells from mouse bone marrow. *J. Immunol. Methods* **223**, 77–92 (1999).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank C. Eigsti and K. Holmes of the Flow Cytometry Unit for FACS sorting, S. Cooper for help with the mouse care, S. Hieny for providing *Toxoplasma gondii* tachyzoites, M. Udey for helpful discussion, and A. Sher and G. Milon for critical reading of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.B. (e-mail: yasmine.belkaid@cchmc.org).

We've got mice...

A new zebrafish microchip — but this week it has to be mainly mice.

Mouse Universal Reference Total RNA

Clontech www.clontech.com

Raising the standards

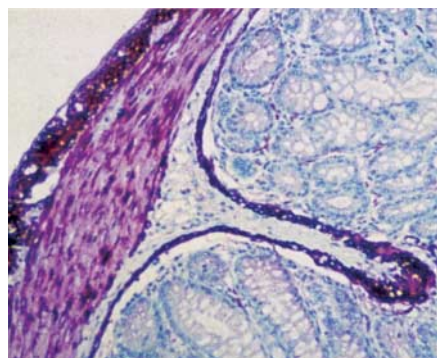
Clontech claims that comparisons of your microarray data just got easier with the introduction of Mouse Universal Reference Total RNA. This universal reference RNA is made by pooling the total RNA extracts from a collection of different tissues, yielding a mixture with the broadest possible gene representation available. Production is on an industrial scale, which minimizes variation between lots. Each kit is supplied with sufficient RNA for up to 80 microarray experiments, facilitating comparison of data sets from different microarray experiments. It can be used with any standard array or labelling method.

IHC kit for rodent tissues

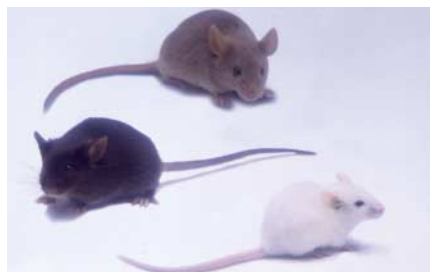
InnoGenex www.innogenex.com

Close encounters

This kit is designed for immunohistochemical (IHC) staining of tissues from species closely related to the primary antibody-producing species. The biotinylated secondary antibodies have been treated with solid-phase adsorption against protein extracts from mouse, rat or rabbit tissues to eliminate cross-reactivities and produce clean results with low background. Serial amplifications allow high sensitivity, retaining the high specificity inherent in the antibody-antigen and biotin-streptavidin interactions. The kits are available for mouse antibodies on rat tissues, rabbit antibodies on mouse or rat tissues, rat antibodies on mouse tissues and mouse antibodies on rabbit tissues, with a choice of AEC, BCIP/NBT, DAB, Fast Red, Phycoerythrin, FluorBlue, FluorGreen or FluorRed chromogen/substrate. Each kit contains sufficient reagents for staining 60 slides. Results are obtained in about two hours.



Staining made easy: IHC kits from InnoGenex.



Focus on tau proteins with Taconic.

Mouse model for neurological disorders

Taconic www.taconic.com

It takes tau to tangle

Taconic's tau micro-injected mouse model carries the transgene for the human P301L mutation of the microtubule-associated protein tau (*MAPT*) gene. Fronto-temporal dementia and parkinsonism linked to chromosome 17 have been correlated with mutations in the *MAPT* gene encoding the tau protein. Motor and behavioural disturbances are observed in approximately 90% of these mice by 10 months of age, correlating with the development of neurofibrillary tangles. The tau micro-injected mouse is suitable for *in vivo* research studies of Alzheimer's disease, Pick's disease and other syndromes associated with neurofibrillary tangles.

Zebrafish 14K array

MWG Biotech www.THE-MWG.com

For good eggs

Described as the first commercially available biochip on this important model organism, the zebrafish array contains 14,067 zebrafish genes. Each oligonucleotide represents a protein-coding sequence of the zebrafish genome. The array is designed to be used in embryogenesis research. The zebrafish, as a vertebrate, is especially suitable for this type

of research as its eggs are translucent and mature outside the mother's body.

HistoMouse-MAX

Zymed www.zymed.com

Mouse-to-mouse enquiries

HistoMouse-MAX can be used to detect mouse, rabbit, rat and guinea pig primary antibodies on mouse or rat tissue samples. It contains no biotin or streptavidin, so potential background due to endogenous biotin activity is avoided. Zymed's BEAT (blocking endogenous antibody technology) prevents the secondary antibody from reacting with endogenous Ig. HistoMouse-MAX contains high sensitive peroxidase (HRP) conjugate and uses one step of HRP conjugated to second antibody instead of two steps as in the LAB-SA method. Kits are available in the HRP/DAB system, producing a dark brown stain, and the HRP/AEC system, which creates an intense red.

AnonyMOUSE

Mouse Specifics www.mousespecifics.com

Non-invasive cardiac screening system

AnonyMOUSE is a high-throughput system for recording electrocardiograms (ECGs) in conscious mice without the use of anaesthesia, restraint, surgery or implants. The system detects the electrical signals of the heart through the feet of small rodents, providing heart-rate measurements and showing any evidence of cardiac arrhythmias. AnonyMOUSE has been used successfully to demonstrate ECG properties in dystrophin-deficient mice that mirror clinical observations in patients with Duchenne muscular dystrophy.

These notes are compiled in the Nature office from information provided by the manufacturers.



nature

the mouse genome

Human biology by proxy



Cover illustration
by Leslie Gaffney and Eric Lander, created by Runaway Technology Inc. using PhotoMosaic by Robert Silvers. It is used courtesy of the Whitehead Institute for Biomedical Research.

Since the publication of the human genome, the scientific community has been eagerly awaiting the results of the mouse genome sequencing project. This week's issue contains a landmark publication from the Mouse Genome Sequencing Consortium that many say holds more promise for our future than even the human genome itself. But why? The laboratory mouse is hailed as holding the experimental key to the human genome. Working on mouse models allows the manipulation of each and every gene to determine their functions, and this will give us detailed insights into many aspects of human disease as well as basic human biology. It is our pleasure to present this special section of *Nature* on the mouse genome. We hope it will provide a comprehensive overview of the future of mouse, and by extension human, genetics and genomics.

The 2.5-Gb mouse genome sequence reported on page 520, from the C57BL/6J strain, reveals about 30,000 genes, with 99% having direct counterparts in humans. To further understand physical gene structures, a second consortium describes in a companion paper 37,086 individual 'transcriptional units' (page 563). This cDNA resource is freely available as physical clones and will advance the pace of research. Humans appear to have about the same number of genes, with similar sequence, and we both like cheese. So why aren't mice more like us? The answer probably lies in the regulation of those genes.

In a third paper, Dermitzakis *et al.* directly compare sequences on human chromosome 21 with their counterparts in mice (page 578). Surprisingly, even gene-poor regions of the chromosome show extensive similarity between the two organisms, suggesting that further exploration into the non-coding regions of both genomes will be required to explain the differences between us. In a separate study, Wade *et al.* compare two inbred mouse strains to map regions of high and low polymorphism (page 574). Distinguishing areas of high polymorphism will be extremely useful for mapping quantitative trait loci, and thus possibly identifying genes that contribute to the inheritance of complex diseases. Finally, in an effort to gain functional information, two groups report large-scale analyses of gene expression in both adult and embryonic tissues of mouse orthologues of the genes on human chromosome 21 (pages 582 and 586). The expression patterns of these genes provide a first step in generating information on specific gene function.

As the sequence has been deposited in public databases, the mouse genome and resources included here provide a huge boost to genetic researchers. Given the bonanza of information they have just received, scientists may now consider the laboratory mouse to be 'man's best friend'.

Chris Gunter Associate Biology Editor
Ritu Dhand Chief Biology Editor

timeline

510 The mouse genome

commentary

512 Mining the mouse genome

news and views

515 Comparative genomics:
The mouse that roared

517 Single nucleotide
polymorphisms: Tackling
complexity

518 Functional genomics:
A time and place for
every gene

articles

520 Initial sequencing and
comparative analysis of
the mouse genome

563 Analysis of the mouse
transcriptome based on
functional annotation of
60,770 full-length cDNAs

letters to nature

574 The mosaic structure of
variation in the laboratory
mouse genome

578 Numerous potentially
functional but non-genic
conserved sequences on
human chromosome 21

582 Human chromosome 21 gene
expression atlas in the mouse

586 A gene expression map of
human chromosome 21
orthologues in the mouse

timeline

The house mouse, *Mus musculus*, has been inextricably linked with humans since the beginning of civilization — wherever farmed food was stored, mice would be found. Many of the advances in twentieth-century biology owe a huge debt to the mouse, which has become the favoured model animal in most spheres of research. With the completion of the draft sequence of its genome published in this issue, the mouse promises to continue to provide us with an essential resource for all aspects of biology. In this timeline, we chart the key events in the history of the mouse that led to this landmark achievement.

75–125 million years ago A common ancestor of mice and humans



The common ancestor of mice and humans — a creature about the size of a small rat — lives alongside the dinosaurs. The picture above is 125-million-year-old *Eomaia scansoria*, the earliest known representative of the Eutheria lineage, which gave rise to all modern placental mammals.

~6 million years ago *Mus*



The genus *Mus* is established, although the name comes much later, via Latin, from the ancient Sanskrit word 'mush', meaning to steal. *Mus musculus*, the house mouse, does not appear as a distinct species until after the end of the last ice age, at about 8,000 BC. The picture shows an early depiction of the mouse in a detail from a Chinese zodiac dating from AD 877.

1900 Fancy mice become lab mice

Generations of 'fancy mice' are created during the nineteenth century, as hobbyists selectively breed the

house mouse.

These mice have a range of different coat colours, and among the mutations created are agouti and satin, both still used in research today. In 1900, Abbie Lathrop, a retired schoolteacher, turns her hobby of breeding fancy mice into a business, selling the animals from her farm in Granby, Massachusetts. Within two years, biologist William Ernest Castle (above) has begun buying mice from her to conduct experiments in his lab in nearby Harvard, testing Mendel's laws of inheritance on coat colours.



1909 Birth of the lab mouse



Clarence Cook Little (above), a Harvard biologist from William Castle's lab, develops the first inbred mouse strain, known as DBA (dilute brown non-agouti). He is convinced that studying a genetically pure breed will unlock the secrets of human diseases such as cancer. This event is subsequently hailed as the birth of the modern lab mouse, and DBA mice are still used in genetics labs today.

1921 C57BL

Clarence Little breeds the mouse strain C57BL from a female mouse code-numbered 57 bought from Abbie Lathrop's farm. C57BL becomes one of the most widely used and important mice to geneticists, and is the strain that will have its genome sequence completed and published in 2002.

1929 The Jackson Laboratory



Two car barons, Edsel Ford (son of Henry) and Roscoe Jackson, head of the Hudson Motorcar Company, provide backing for Clarence Little to set up the Jackson Laboratory in Bar Harbor, Maine. The lab grows into one of the world's most important research centres for mouse genetics.

1972 First computer database for mouse genetics

Mouse Genome Informatics

The Jackson Laboratory updates its card-file database for mouse genetics by designing the first computer database for mammalian genetics. It forms the precursor to the Mouse Genome Database, which will serve as the hub of mouse-genome sequencing projects.

1982 First transgenic mouse



A team led by Richard Palmiter and Ralph Brinster fuse elements of a gene that can be regulated by dietary zinc to a rat growth-hormone gene, and inject it into fertilized mouse embryos. The resulting mice, when fed with extra zinc, grow to be huge, and the technique paves the way for a wave of genetic analysis using transgenic mice.

1987–89 The first knockout mouse

Teams led by Martin Evans, Oliver Smithies and Mario Capecchi create the first 'knockout' mice, by selectively disabling a specific target gene in embryonic stem cells. The three receive the Lasker Award in 2001 for this achievement, and the technique goes on to be used to create several thousand knockout mice.

1998 Mouse clones

After Dolly the cloned sheep in 1997, a team in Hawaii produces Cumulina and her clones, the first cloned mice.

JACKSON LAB

JACKSON LAB

JACKSON LAB

M. A. KLINGER/CNH

BRITISH LIBRARY/INT. DUNHUANG PROJECT

The life history of the mouse in genetics

1999 The Mouse Genome Sequencing Consortium



At its inception in 1990, the Human Genome Project included the mouse as one of its five central model organisms to have its genome sequenced. In 1999, with the human genome sequence well under way, three major sequencing centres (the Wellcome Trust Sanger Institute, the Whitehead Center for Genome Research and Washington University Genome Sequencing Center) launch a concerted effort to sequence the mouse genome, under the title of the Mouse Genome Sequencing Consortium. By 2002, the consortium has expanded to embrace scientists from 26 institutes in six countries.

2001 Feb The human genome

The first drafts of the human genome are published. The publicly funded International Human Genome Sequencing Consortium publishes its draft sequence in *Nature*, going head-to-head with Celera Genomics, a US-based company led by Craig Venter, which publishes in *Science*. Both of these milestone achievements provide a panoramic view of our genomic landscape, estimating a



size of 3.2 billion base pairs (Gb) and predicting some 31,000 genes. Much work needs to be done to produce a complete sequence, but the collaborative effort of the public consortium coupled with the Celera sequence reflect an unprecedented achievement.

2001 June The private shotgun mouse genome



Celera sells its draft mouse sequence, which has been generated by whole-genome shotgun sequencing — a technique in which the entire genome is sequenced at once. Five-fold coverage of the mouse genome is made available to private customers for \$45,000 for a three-year contract. Celera's sequence is made up of four

strains of mouse, including DBA, but not C57BL/6J — the subject of the publicly funded sequencing effort.

2002 May Mouse chromosome 16

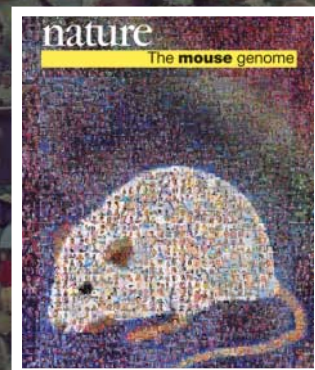
Using the sequence generated by Celera's private mouse genome, Richard Mural *et al.* publish an analysis of the sequence of mouse chromosome 16 in *Science*. They find that there is a large region with high genetic similarity to human chromosome 21, which has already been well characterized.

2002 Aug Physical map of the mouse genome



An international consortium provides the framework for the full determination of the mouse DNA sequence by building a physical map of the genome. Large blocks of correspondence between mouse and human sequences make it possible to lay out chunks of decoded mouse sequence and, conversely, the mouse map can be used to fill gaps in the human genome.

2002 Dec The mouse genome



The Mouse Genome Sequencing Consortium publishes the culmination of international efforts — a high-quality draft sequence and analysis of the genome of the C57BL/6J mouse strain. The estimated size is 2.5 Gb, smaller than the human genome, with less than 30,000 genes. About 40% of the human and mouse genomes can be directly aligned with each other, and about 80% of human genes have one corresponding gene in the mouse genome. Accompanying papers detail various other aspects of the genetic make-up of the mouse, and through the efforts of the RIKEN Genome Science Laboratory in Japan, many essential resources are made freely available.

For an extended, interactive version of this timeline, visit our Mouse Genome Web Focus at www.nature.com/nature/mousegenome, which also contains all of the articles on the mouse genome from this special issue, available free. These are accompanied by related articles from the archives of *Nature* and *Nature Reviews Genetics*, available free until 5 March 2003.

Mining the mouse genome

We have the draft sequence — but how do we unlock its secrets?

Allan Bradley

The mouse genome sequence, published in this issue¹, has already made a huge impact on the research community. Although only a draft, it is clear that the sequence is a very high-quality product, with excellent coverage and reliability over large genomic expanses. It is a huge asset to researchers, and its significance matches that of the human genome. In the past six months, for example, the Ensembl genome browser of the Sanger/European Bioinformatics Institute dealt with 2.6 million requests for detailed information about the mouse genome, and 3.2 million queries about the human sequence.

But there is one important difference between these two resources — the mouse genome encodes an experimentally tractable organism. This means that it is now truly possible to determine the function of each and every component gene by experimental manipulation and evaluation, in the context of the whole organism.

An ideal tool

Over the past two decades, the mouse has emerged as the pre-eminent model organism for two fundamental reasons. First, it is a mammal, and so has many physiological, anatomical and metabolic parallels with humans. Although the anatomical differences between humans and mice appear striking, they reflect alterations in size and shape — detailed analysis of organs, tissues and cells reveals many similarities, extending to whole-organ systems, physiological homeostasis, reproduction, behaviour and disease. The mouse is an excellent surrogate for exploring human biology, and disease processes in the animal can accurately reflect those in humans. This explains why the mouse is widely used to investigate diverse aspects of mammalian biology and pathology, ranging from embryonic development to metabolic disease, behaviour and cancer in adults.

Second, although it is certainly true that mammalian biology is available in other species that in some cases are closer to humans, the mouse has one feature that has been uniquely developed compared with all other mammals — genetic tractability. The similarities in biology and pathology between mouse and human are reflected in the genomes. For virtually every gene in the human genome, a counterpart can readily be identified in the mouse. Genetic manipulation within the living mouse is routine and can these days be done with extraordinary precision. Consequently, many mouse strains have been generated with genetic lesions that echo those observed in human genetic disease.

In some cases, these manipulations yield a model that closely resembles the pathology of the analogous human condition — for example, the susceptibility to cancer of mice deficient in the gene encoding the protein p53 resembles that of humans who have mutations of their p53 gene. In other cases, only some aspects of the human pathology are apparent; for instance, mice with mutations in the cystic fibrosis gene do not develop lung disease, which is the most devastating aspect of the condition in humans. Understanding how the disease process is suppressed in mice will provide important clues for treating the human condition. Genetic studies in mice are greatly helped by the availability

of inbred strains, which allow experimental parameters to be measured in a homogeneous genetic background.

The mouse genome sequence¹ is already fuelling the next phase of research. At a very fundamental level, we now have a reasonable 'parts list' for the mouse. Some of these parts — the transcription units — are the elements we understand best. Although there are now many new gene transcripts to explore, existing experimental methods can be used to examine them. But the mouse genome also contains many non-coding regions of sequence identical to their counterparts in the human genome. New experimental methods will have to be deployed to examine the function of these conserved sequences.

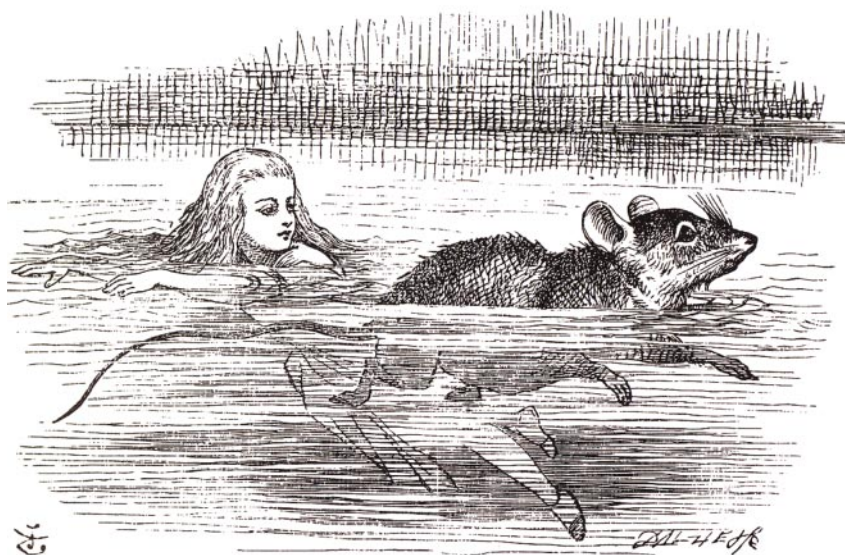
The complementary DNA (cDNA) clone and sequence resource described on page 563 by a group from the Institute of Physical and Chemical Research (RIKEN) in Japan², illustrates the potential usefulness of the mouse genome sequence. The authors' comparison of their FANTOM clones with the genome sequence immediately illustrates gene structures that will enable the mutagenesis work discussed below. The physical cDNA resource can be used in many experimental situations: for instance, in over-expression studies in cell lines or transgenic mice (in which the product of the gene concerned is made in abundance to allow its function to be investigated); to produce proteins for structural studies or antigens to obtain antibodies for investigating gene expression; or in studies of interactions between different proteins. For laboratories interested in a single cDNA clone, knowledge of the end-points of the messenger RNA is sufficient to design primers to rapidly retrieve a cDNA by reverse-transcriptase PCR (polymerase chain reaction) analysis, saving months of time 'walking' through cDNA libraries. The mouse cDNA parts list is already being used to illuminate patterns of gene expression.

Although the draft mouse genome sequence is available now, it will be two years before it is a finished, reference-quality product. Even so, the annotation of function to genes is already under way on a small scale in many laboratories, and larger-scale studies are being initiated. Unlike DNA sequencing, experimental work is not always amenable to high-throughput, automated approaches and can be complex to interpret. The research community clearly faces an enormous task — one that will extend well beyond the next decade.

Joined-up efforts

The assembled mouse genome provides a framework onto which functional information will increasingly be layered. Genome browsers (see Box, overleaf) provide an interactive graphical view of the genome, rendering its vast size (equivalent to a million pages of text) accessible to the scientific community. They can provide detailed information about a single gene, or enable a wider perspective of the genomic landscape in a single species or across several species.

Mutation data, phenotype information and gene-expression data (described below) are just three of the many potential data sets that have to become accessible by these means. Many laboratories generating large data sets with 'post-sequence' goals in mind recognize the need to make their data accessible electronically. But few have sought ways to link their data back to a browser, and many may not have



Pooled resources: the mouse sequence will offer insight into the workings of the human genome.

access to the computational infrastructure needed to respond to a large volume of queries. The mouse genetics community has enjoyed an exceptionally high-quality mouse genome database for many years (see Box, overleaf), but this must continue to evolve and be intimately integrated with genome browsers. Joined-up data sets are essential if we are to reap the potential rewards from the mouse genome.

The draft mouse genome is being intensively scrutinized, but how much functional information can be deciphered from sequence-gazing alone? Which experimental approaches will provide the greatest information for the resources spent? Computational prediction and alignment programs can already identify many genes and predict some aspects of gene structure. Yet computers remain inferior to cellular machinery in recognizing a gene, where it starts and stops, and when it should be turned off and on. At best, computer predictions provide a very incomplete picture of a genome. Detailed 'hand-crafted' curation, coupled with experimental analysis, is necessary to clarify the encoded gene set. The paper from RIKEN² goes some way towards satisfying the need to identify the complete set of mouse genes. This resource must be distributed widely and quickly, without restrictions on access or usage, if the value of the clone set is to be realized.

Untangling the genome

Computational analysis is also being used to classify genes into families based on conserved motifs in the gene sequences, although such functional insights are limited to some classification of a protein's biochemical activity or cellular function based on the motifs. For instance, transcription factors and enzymes can be recognized from their motifs, but not

Joined-up data sets are essential if we are to reap the potential rewards from the mouse genome.

where and when they would be expressed and who their partners are. This information is encoded in the genome, but it is indecipherable at present and so impossible to extrapolate to the physiological role of any gene.

Considerably more experimental information is needed to predict gene function, but which experimental approaches are applicable to high-throughput analysis of large gene sets in complex multicellular organisms? Several groups believe that information on gene expression is invaluable and should be generated for every gene in the genome. The expression pattern of a gene in a multicellular organism is a basic feature of the biological function of any gene, whatever its function. The more contexts in which expression is examined, the greater the insight into function. In principle, a definitive and comprehensive atlas of the expression pattern of every gene in the genome can be generated.

One experimental approach in which thousands of genes can be analysed in parallel is to isolate messenger RNA and to display the gene-expression profile on a chip. When this technique is applied to tissues, data are lost because aspects of the three-dimensional structures of multiple cell types are destroyed in the biochemical extraction. Data from *in situ* analyses contain more detailed information about each gene, but the generation of these data is serial and significantly slower.

Gene expression is being systematically examined at the transcriptional level by several groups, for instance in the 9.5-day-old mouse embryo and in adult tissues (see Box, overleaf). Two other papers in this issue^{3,4} report large-scale analyses of gene expression in embryonic and adult stages, but so far have examined just 0.5% of the genes in the genome, the homologues of the genes on chromosome 21. Transcription studies *in situ* have relatively limited resolution, and the tissues constituting a multicellular organism are complex mixtures of different cell types. Unless each cell is individually visualized for gene expression in combination with histological criteria, important information relating to biological function is lost, for instance the subcellular compartment(s) occupied by a protein.

The Sanger Institute's Atlas project is being established to systematically examine the expression pattern of every gene product at tissue-, cellular- and subcellular-level resolution, to provide a permanent, definitive and accessible record of the molecular architecture of normal tissues and cells. The ultimate goal is to define protein expression patterns for all 30,000 mouse genes in hundreds of different tissues, all gathered in archival data sets to support research projects worldwide. Data will be collected electronically and archived with a vocabulary allowing complex queries.

A protein's location in a cell and a tissue provides important clues about its function, but such data are still insufficient to reveal its physiological role *in vivo*, or the temporal and spatial specificity of gene products for an as-yet-unknown functional activity. Gene-expression data are informative and can guide an experimental path, but cannot be interpreted in isolation.

Mutational analysis

One of the most informative experimental approaches for examining gene function is to analyse mutants. Spontaneous and induced mutations in the mouse have been studied for more than 100 years, but in the past decade there has been an explosion in their use.

The isolation of embryonic stem (ES) cells⁵ and demonstration that these cultured cells can recolonize the mouse germ line⁶ were the two fundamental discoveries that led to the first 'knockout' mouse in 1987, through a genetic modification that had been engineered *in vitro*. This heralded a golden era for mouse genetics. Today, it is possible to engineer mice with genetic changes as subtle as a single nucleotide substitution or with major alterations of the genome such as the deletion or duplication of millions of base pairs⁷. The genetic tractability of ES cells has made the mouse uniquely accessible for genetic studies compared with every other multicellular organism.

Despite this success, the combined output of the mouse genetics community over

the past 10 years has described mutations in just a few thousand genes by this method, 10–15% of the predicted gene content of the organism. Can this rate be speeded up so that it will not take 50 years to mutate and analyse the remaining 85–90%? Several leading mouse genetics laboratories have begun to discuss plans to generate a knockout for every gene in the mouse genome, but how will this be achieved?

Gene targeting requires detailed knowledge of gene structure to ensure that the target locus has been effectively mutated. One of the most immediate practical benefits of the assembled mouse genome sequence is the availability of detailed gene-structure information, enabling mutations to be made with a full understanding of the likely functional consequences. Knowledge of the genome sequence has also made it possible to index libraries of gene-targeting vectors, eliminating the need to screen a library to obtain a genomic clone for targeting. Library indexing by end-sequencing significantly increases the rate at which knockout mice can be generated.

On target

Although the genome sequence has enhanced the rate at which targeted mutations can be generated, is this going to be fast enough? Targeting is an inherently serial process — a single experiment typically generates one type of allele. Gene trapping, on the other hand, in which genes are tagged for sequence retrieval by insertional mutagenesis, generates hundreds of different mutations from a single electroporation or viral infection. Recognizing the importance of a genome-wide gene-trap library, but unable to fund this in the academic sector, I and other colleagues set up Lexicon Genetics. Although the Lexicon resource is now quite comprehensive, the cost is significant, intellectual property rights may have to be negotiated, and some mutants will not be available for commercial reasons.

There are also gene-trap libraries in the public sector (see Box), through which about 16,000 ES cell clones are now available. In principle, these resources should be distributed with few constraints. Although their coverage is more limited and the effort less centralized than the Lexicon library, over the next couple of years this resource should expand considerably. The value of such an archive will be fully realized only if it can be exploited by the community. One key aspect of the elaboration of this resource is to ensure that the knowledge of the location of these mutations in the genome is linked with access to the physical resource (the trapped ES cell clone), so that ES cell clones can be retrieved and used to establish mice carrying these mutations.

Over the past decade, knocking out genes has provided a rich source of information

Web links

Genome browsers

- ▶ www.ensembl.org
- ▶ www.ncbi.nih.gov/genome/guide/mouse
- ▶ genome.ucsc.edu

Expression information

- ▶ genex.hgu.mrc.ac.uk
- ▶ bodymap.ims.u-tokyo.ac.jp
- ▶ tigem.it
- ▶ chr21.molgen.mpg.de/data

ENU mutagenesis centres

- ▶ www.mouse-genome.bcm.tmc.edu
- ▶ www.mgu.har.mrc.ac.uk/mutabase
- ▶ pga.jax.org
- ▶ www.jax.org/nmf
- ▶ www.gsf.de/ieg/groups/enu-mouse.html
- ▶ jcsmr.anu.edu.au/group_pages/mgc/MedGenCen.html
- ▶ cmhd.mshri.on.ca
- ▶ tnmouse.org

Gene-trap consortium sites

- ▶ www.genetrap.de
- ▶ www.escells.ca
- ▶ baygenomics.ucsf.edu
- ▶ www.lexgen.com

Mouse genome database

- ▶ www.informatics.jax.org

about gene function — and as a result, this sequence-driven approach has been widely adopted. But there is considerable uncertainty in predicting the phenotypes that will be displayed by the mutant mice. In my own view, selection of a candidate gene for mutational analysis might as well be stochastic as based on assimilation of existing knowledge.

Sadly, our knowledge of conserved domains, expression patterns, biochemical activity, protein–protein interactions and molecular structure is inadequate to predict function. A knockout phenotype often shamelessly displays our collective ignorance about gene function. We have to accept that in many cases, sequence-directed mutagenesis may not efficiently identify genes specific to certain functions or disease — for instance, the genes involved in diabetes — by knocking out individual candidates.

Mutational analysis can be focused to identify the players in a specific process by performing a genetic screen, as widely used in organisms such as the yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, and the fruitfly *Drosophila melanogaster*. Screens did not catch the imagination of the mouse community when first introduced 15 years ago because they emerged in parallel with gene-targeting approaches, which looked so promising. Over the past few years, genetic screens have become much more popular because of a few

high-profile successes⁸. Currently, ENU (*N*-ethyl-*N*-nitrosourea) mutagenesis is the most widely used method for random mutagenesis, and several programmes using this approach have been initiated (see Box).

Although the 1,000-plus new mutations generated by ENU mutagenesis provide a resource for future studies, they will not add any knowledge about gene function until the underlying genetic lesions have been identified, and the molecular mechanisms relating the lesion to the observed phenotype are understood in detail. This leaves the strategy with two major bottlenecks — the identification of the mutation (typically a nucleotide substitution); and a detailed phenotypic understanding of each mutant.

Community work

Whatever method is used to generate a mutation, understanding the mechanistic cause of the observed phenotype is central to determining gene function. This understanding depends not only on knowing the mutated gene, but also on having a very detailed picture of the phenotype. Several centres are developing standardized expertise in this area, but high-throughput screens will not detect subtle variations from normal, nor will they examine mice for every possible phenotype. Some of the most valuable screens will be pursued in the context of very detailed phenotyping, possibly in the context of another genetic alteration in the background of the mice being screened.

So the job of phenotyping mutants for a specific characteristic is a task that cannot be easily delegated to a centre; rather, it is an activity for the whole community. This requires the mobility of existing strains between groups and availability of funding to pursue smaller, more scientifically focused, screens in specific areas of biological expertise. Progress in understanding the mouse genome will involve the input of diverse experimental approaches in thousands of small and a few large laboratories. The accessibility of mutants is key to this progress, and this comes with a cost — because strains will need to be maintained or archived for decades. So the avalanche of genome sequence will be followed by an explosion of mutant mice, requiring new mouse facilities to house and phenotypically evaluate this global genetic resource. ■

Allan Bradley is at the Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

1. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
2. The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team *Nature* **420**, 563–573 (2002).
3. Raymond, A. et al. *Nature* **420**, 582–586 (2002).
4. The HSA21 Expression Map Initiative *Nature* **420**, 586–590 (2002).
5. Evans, M. J. & Kaufman, M. H. *Nature* **292**, 154–156 (1981).
6. Bradley, A., Evans, M. J., Kaufman, M. H. & Robertson, E. J. *Nature* **309**, 255–256 (1984).
7. Ramirez-Solis, R., Liu, P. & Bradley, A. *Nature* **378**, 720–724 (1995).
8. Vitaterna, M. H. et al. *Science* **264**, 719–725 (1994).

The mouse that roared

Mark S. Boguski

The laboratory mouse has become an indispensable tool for investigators in many areas of biomedical research. The availability of the full mouse genome sequence will immeasurably advance both the character and the pace of discovery.

"Know then thyself, presume not God to scan; The proper study of mankind is man," wrote Alexander Pope in 1733. What better reason could there have been to sequence the human genome? But the planners of the Human Genome Project realized that the data could not be fully understood, or used to advance biomedicine, in isolation. Indeed, many of the "lessons learned and promises kept"¹ have been derived from the study of model organisms. *Mus musculus*, a species of mouse, has been one of the five key model organisms sequenced since the beginnings of the Human Genome Project. In 1998–99 the US National Institutes of Health published an action plan for mouse genomics² which, among other things, called for a working draft sequence of the mouse genome by 2003. An international Mouse Genome Sequencing Consortium³ has now achieved this goal. The particular mouse strain concerned is C57BL/6J (Fig. 1), and the consortium's findings are reported on page 520 of this issue.

Why is this so important? It is because there can scarcely be a major area of mammalian biology or medicine to which mouse studies have not contributed in some way, often as surrogates for human studies. For genetics and development, for immunology and pharmacology, for cancer and heart disease, even for behaviour, learning and memory and psychiatric disorders^{4,5}, the laboratory mouse has become an indispensable tool. Much of this power has come from technologies to manipulate the mouse genome, but until now we have in effect been shooting in the dark. The genome of *Mus musculus* will provide the necessary illumination.

Most of the findings reported by the sequencing consortium³ were revealed by computational sequence analysis, the overall approach being known as comparative genomics^{6–8} when performed on the genome scale. Genetic sequences — strings of nucleotide bases — are "documents of evolutionary history"⁹, from which much information can be inferred from their conservation or divergence and rearrangement relative to a common ancestor. Genome analysts have applied this approach at many levels, from multi-megabase rearrangements reflected in chromosome structure down to single nucleotide changes between orthologous genes — two genes are orthologous if they diverged after a speciation event, when a new species forms from an existing one; two genes are paralogous if they diverged after a gene duplication event. One of the outcomes of a comparative genomic analysis is an

enumeration of the total number of genes shared by two species (see mammalian gene count, below).

An unexpected finding to emerge from the sequence of the human genome is that we appear to have far fewer protein-coding genes than previously suspected — fewer than 30,000, compared with the figure of 80,000–100,000 frequently cited in textbooks published before 2001. Analysis of the mouse genome backs up this finding. The sequencing consortium estimates that it contains 27,000–30,500 protein-coding genes. Ninety-nine per cent of these genes have a sequence match in the human genome and 96% of these lie within 'syntenic' regions of mouse and human chromosomes.

The consortium was able to align long segments of mouse and human chromosomes in which the linear orders of genes in the common ancestral sequences were conserved. Although this so-called 'conservation of synteny' between mouse and human chromosomes has long been recognized, the comprehensiveness and precision afforded by the genome sequences will allow effective cross-reference of the locations of any genetically mapped traits in the mouse with genes in the orthologous regions of the human genome (and vice versa). This will greatly accelerate the isolation of disease genes. It will also be important for precise deletion (knockout) of mouse genes to study their functions and for targeting human sequences to their syntenic locations in the mouse genome, allowing the mice to be 'humanized' for various traits.

Based on pairwise alignments of nearly 13,000 (out of about 28,000) orthologous gene pairs, the consortium found that the encoded proteins had a median amino-acid sequence identity of 78.5%. In comparison, orthologous mouse and rat proteins are, on average, 97% identical¹⁰, and a sample of human and *Caenorhabditis elegans* (nematode)

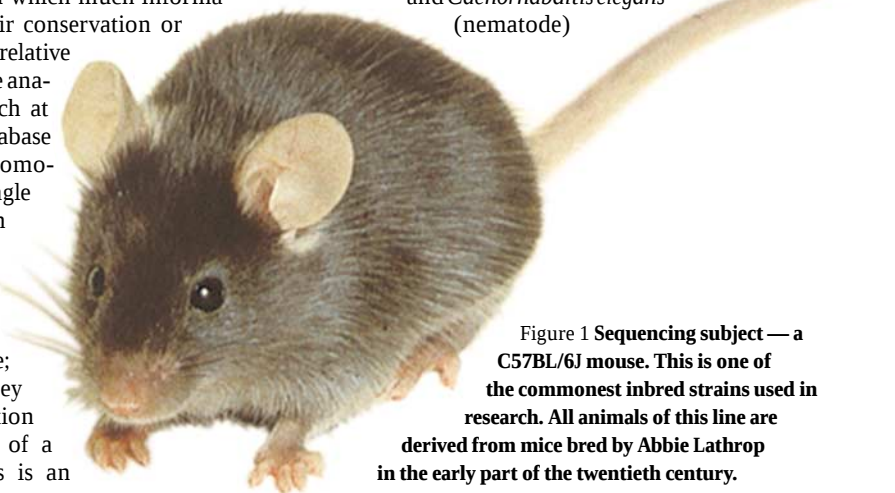


Figure 1 Sequencing subject — a C57BL/6J mouse. This is one of the commonest inbred strains used in research. All animals of this line are derived from mice bred by Abbie Lathrop in the early part of the twentieth century.

proteins had an average of 49% of their amino acids in common¹¹. For many of the unaligned human and mouse genes³, it was not possible to confidently identify one-to-one orthologous pairs because of differential gene expansion — contraction or loss in various gene families that results in one-to-many or even many-to-many pairings. It is undoubtedly in the divergent evolution of these gene families that some of the differences between humans and rodents will be found.

There is a highly practical aspect to this kind of study. For example, the consortium describes the expansion of four subfamilies of cytochrome P450 (CYP) genes in mice compared with humans. These genes encode enzymes that function in drug metabolism, and biologically are part of a general xenobiotic 'sensory system' for chemical compounds in the environment. In the pharmaceutical industry, one stage of drug development is known as 'ADME/Tox' — shorthand for studies of the absorption, distribution, metabolism, excretion and toxicology of drug candidates. Mice and humans have been exposed to different types and amounts of environmental substances (including toxins) during their evolutionary divergence, and one might expect natural selection to have altered the numbers and activities of these enzymes in the two species. Indeed, humans and rodents do often have different responses to drugs and other chemicals. However, using transgenic technologies — which will be improved with the availability of the genome — mice can be 'humanized' to metabolize drugs more as we do¹².

The free availability of mouse sequence data, throughout the project, has already allowed the identification of novel genes through comparative genomics. For instance, a gene called *APOAV* encodes a previously unknown member of the apolipoprotein gene family. A remarkable aspect of this story is that the last member of this gene family, *APOAIV*, was cloned nearly two decades ago¹³ in the now well-trodden fields of lipid metabolism and cardiovascular disease. After years of thinking that all of the members had been catalogued, the discovery of *APOAV* was quite surprising¹⁴.

The *APOAIV*, *APOCIII* and *APOAI* genes occur within a 20-kilobase stretch of DNA on human chromosome 11 (ref. 15). Through comparative analysis of human and mouse genomic sequences, Pennacchio *et al.*¹⁴ identified a region of conserved sequence, some 25 kilobases from *APOAIV*, that proved to contain the *APOAV* gene. Because the *AIV/CIII/AI* gene locus was known to influence plasma lipid levels in humans, Pennacchio *et al.* studied lipid levels in knockout and transgenic mice and found that the *APOAV* protein has a strong inverse correlation with plasma triglyceride levels — a risk factor for coronary artery disease.

Box 1 Sequencing priorities

Much genome-sequencing capacity is still devoted to finishing the human, mouse and rat genomes. But as that work is completed, laboratories will become free to take on new projects. The US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland, one of the main funding agencies for sequencing, has called for 'white paper' proposals for future priorities among candidate organisms. One stipulation for NHGRI support is that data should be released rapidly and without restriction.

The first review of proposals by an expert panel, announced in May of this year, gave high priority to sequencing the genomes of the chicken, chimpanzee and honey bee, as well as several species of fungi, a sea urchin and a protozoan, *Tetrahymena thermophila*. Sequencing work on plants and the Bacteria

and Archaea is supported by different agencies.

In a second announcement, in September, the cow and dog were made high priorities, along with *Oxytricha*, a single-celled organism that is of especial interest to geneticists and evolutionary biologists. The cow is evidently of great importance in agriculture, whereas dogs provide good models for certain human diseases, such as narcolepsy and obsessive-compulsive disorder. Along with a sequenced genome of the chimpanzee, the data on both animals will also provide grist for comparative studies among mammals.

Among the criteria for selection are the likely demand for the sequence data from particular research communities, the genome size and projected cost of the project, and whether a complete sequence is necessary.

Subsequent studies¹⁶ showed that genetic variation in the *APOAV* locus influences plasma triglyceride levels in humans.

This example is but one harbinger of the impact that the mouse genome will have on human biology, and more systematic efforts will undoubtedly be made to cross-reference known and novel genes in humans and rodents with their biological effects. For instance, the Alliance for Cellular Signaling¹⁷ aims to identify all of the proteins that are involved in signalling pathways in B lymphocytes (antibody-producing cells) and cardiac myocytes (heart muscle cells) in the mouse. By cataloguing all of the orthologous proteins encoded by the human genome, one will eventually be able to move almost effortlessly back and forth between clinical observations in humans and experiments with mouse models of disease. For physiological and pharmacological studies, the rat (not the mouse) has been the long-standing model organism, primarily because of its larger size. A working draft of the rat genome has just become available¹⁸. A proposed 'triangulation' strategy¹⁹ should powerfully leverage the advantages of all three organisms (mice, rats and humans) for studies of human disease.

Finally, what might the mouse genome teach us about mammalian biology and evolution? About two years ago, an article reporting results of the sequencing and analysis of the *Drosophila melanogaster* (fruitfly) genome was boldly entitled "Comparative genomics of the eukaryotes"²⁰ — eukaryotes, loosely, being organisms whose cells have nuclei, in contrast to the Bacteria and Archaea. This was despite the fact that genome-scale data for only four eukaryotes (*Drosophila*, *C. elegans*, the yeast *Saccharomyces cerevisiae* and human) out of several million eukaryotic species on the planet were then available. We must take care not to over-generalize from small sample

sizes. Likewise, analyses of the genomes of humans and mice have led to the notion of a 'mammalian gene count' (currently standing at about 28,000) derived from only two out of an estimated 4,600–4,800 extant mammalian species on Earth. But energetic programmes²¹ to obtain more generally representative data are under way (Box 1). These are likely to deliver an ultimately more satisfying picture of what the late Stephen Jay Gould called the "full house" of biological variation, from cabbages to kings²². ■

Mark S. Boguski is at the Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North D4 100, PO Box 19024, Seattle, Washington 98109, USA.

e-mail: mboguski@fhcrc.org

1. Tilghman, S. M. *Genome Res.* **6**, 773–780 (1996).
2. Battey, J., Jordan, E., Cox, D. & Dove, W. *Nature Genet.* **21**, 73–75 (1999).
3. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
4. Tarantino, L. M. & Bucan, M. *Hum. Mol. Genet.* **9**, 953–965 (2000).
5. Bucan, M. & Abel, T. *Nature Rev. Genet.* **3**, 114–123 (2002).
6. Tugendreich, S., Boguski, M. S., Seldin, M. S. & Hieter, P. *Proc. Natl Acad. Sci. USA* **90**, 10031–10035 (1993).
7. Bassett, D. E. Jr *et al.* *Trends Genet.* **11**, 372–373 (1995).
8. Womack, J. E. & Kata, S. R. *Curr. Opin. Genet. Dev.* **5**, 725–733 (1995).
9. Zuckerkandl, E. & Pauling, L. *J. Theor. Biol.* **8**, 357–366 (1965).
10. Makalowski, W. & Boguski, M. S. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
11. Wheelan, S. J., Boguski, M. S., Duret, L. & Makalowski, W. *Gene* **238**, 163–170 (1999).
12. Xie, W. & Evans, R. M. *Drug Discovery Today* **7**, 509–515 (2002).
13. Boguski, M. S., Elshourbagy, N., Taylor, J. M. & Gordon, J. I. *Proc. Natl Acad. Sci. USA* **81**, 5021–5025 (1984).
14. Pennacchio, L. A. *et al.* *Science* **294**, 169–173 (2001).
15. Boguski, M. S., Birkenmeier, E. H., Elshourbagy, N. A., Taylor, J. M. & Gordon, J. I. *J. Biol. Chem.* **261**, 6398–6407 (1986).
16. Talmud, P. J. *et al.* *Hum. Mol. Genet.* **11**, 3039–3046 (2002).
17. www.afcs.org
18. www.hgsc.bcm.tmc.edu/projects/rat
19. Jacob, H. J. & Kwitek, A. E. *Nature Rev. Genet.* **3**, 33–42 (2002).
20. Rubin, G. M. *et al.* *Science* **287**, 2204–2215 (2000).
21. <http://www.genome.gov/page.cfm?pageID=10002154>
22. Gould, S. J. *Full House: The Spread of Excellence from Plato To Darwin* (Harmony Books, New York, 1996).

Single nucleotide polymorphisms

Tackling complexity

Joseph H. Nadeau

Many traits, including susceptibilities to some diseases, are under complex genetic control. A new way of analysing the mouse genome will be a great help in understanding the interactions involved.

Like Passepartout with Phileas Fogg in their 80-day travels, mice have tagged along with humans in a 10,000-year journey around the world. But these animals have been more than fellow travellers — along the way, new types of mice with unusual features were occasionally found. These genetic variants were collected and much prized, whether for their unique coat colours or patterns, waltzing behaviour, small size or other unusual characteristics. Keeping and breeding such fancy mice has long been a hobby with international appeal. In the eighteenth century, during the Edo period in Japan, the book *Chingan Sodategusa* was published as a guide to mouse breeding. And more recently, around the beginning of the twentieth century, clubs set up by mouse fanciers proliferated in North America and Europe¹.

The transition of fancy mice to laboratory mice began in 1909, with the recognition by C. C. Little that these rich collections, especially when taken as genetically homogeneous inbred strains, could become a powerful resource for biomedical research². Twenty years later, Little (Fig. 1) founded the Jackson Laboratory in Bar Harbor, Maine, and subsequent work there and elsewhere has shown that the mouse is genetically closer to us than had been imagined^{3–5}. Hence the value of the laboratory mouse both for basic research and as a model of human disease, and the significance of the report by Wade *et al.*⁶ (page 574 of this issue), which extends the prospects for detailed investigations of mouse genetics.

The characteristics of some fancy mice result from genes operating in a mendelian fashion; this is the simplest form of inheritance, in which a trait is controlled by two variants, or alleles, of a single gene. Most traits, however, have a genetically complex background and are the subject of considerable attention. For instance, the Mouse Phenome Project is a community effort to characterize the phenotypes — the morphological, physiological and behavioural traits — of inbred strains of mice^{7,8}. The project, which aims to be both comprehensive and systematic, is revealing remarkable phenotypic variability, most of which is probably controlled by numerous genes. Identifying the genes concerned will not only provide insight into fundamental biology, but also new ways of understanding genetically complex human diseases — obesity,

hypertension, diabetes and cardiovascular diseases, for example, as well as drug addiction and the biological underpinnings of learning and memory.

Wade *et al.*⁶ have pioneered a new form of gene mapping in mice that will greatly accelerate gene discovery, especially for genes that control complex traits. Their approach is based on single nucleotide polymorphisms (SNPs) — these are differences of just one nucleotide base in sequences of DNA. Identification of mendelian variants in mice is now routine. But the discovery of genes involved in complex traits is just beginning⁹, and the study of SNPs will help greatly in that endeavour.

The point of departure for Wade *et al.* was the 'finished' DNA sequence of the genome of the C57BL/6J mouse strain. This is the most widely used inbred strain, and the complete sequence of the genome is published formally only now — see page 520 of this issue¹⁰. (Celera Genomics has also produced a genome sequence, as well as a collection of mouse SNPs, but the data have not been deposited in public databases^{5,11}.)

Wade *et al.* compared the C57BL/6J sequence with 59 'finished' segments of the genome of a different (129/Sv) inbred strain,

and found nearly 70,000 SNPs. These SNPs were located in blocks of high SNP density, of 1 million base pairs or more, that were separated by blocks of low variability. The authors estimate that 67% of these genomes are SNP poor, with an average of 0.5 SNPs per 10 kilobases, and 33% are SNP rich, with 40 SNPs per 10 kilobases. There is a sharp transition from high to low SNP density, implying that the difference results from recombination among the chromosomes of the common ancestor^{12,13}, a process in which stretches of DNA are swapped between sister chromosomes.

Wade *et al.* then surveyed a panel of inbred mouse strains and found striking evidence that segments of chromosomes with distinct combinations of SNPs (haplotypes) were shared among common inbred strains. This observation is consistent with both the historical record and genetic evidence from mitochondrial DNA and the Y chromosome that inbred strains have a small number of progenitors^{14–16}. Typically, at any given genomic location, only two or three different haplotypes were found in this panel of strains, suggesting that there is much less genetic variability between them than had been thought. Finally, a survey of strains derived recently from wild mice showed that 67% of each of these genomes are derived from European mice (*Mus musculus domesticus*) and 33% from Asian mice (*M. m. musculus*).

Study of mouse SNPs will accelerate discovery of the genes that underlie complex traits in two ways — by genetic mapping of crosses between phenotypically distinct strains, and by associating a particular



Figure 1 Clarence Cook Little, an innovator in genetics, especially in research with mice.

JACKSON LABORATORY

phenotype with particular genotypes in surveys of inbred strains. The apparently limited genetic heterogeneity among the various strains will facilitate the task of dissecting controls of phenotypic variation. And given that the genomes of humans and mice contain fewer protein-coding genes than had been expected, as Mark Boguski describes in an accompanying News and Views article (page 515), complexity probably arises from the numerous ways in which a modest number of genes and alleles interact, rather than from the mass action of a large number of genes and alleles. Finally, because chromosome segments with high SNP density occupy only about 33% of the genome, the 'search space' for genes controlling complex traits is greatly reduced and the prospects for gene discovery are likewise increased.

Previous technical innovations in research on mouse genetics have delivered striking advances in our understanding. First there was analysis of visible phenotypes; then came study of isoenzyme variation; and since the advent of molecular biology we have had the ability to exploit restriction-fragment-length and simple-sequence-length polymorphisms¹⁷. These innovations facilitated the construction of rudimentary genetic maps, then mapping of sequenced genes and positional cloning of mendelian variants, and now the genetic characterization of complex traits.

We must remember that discoveries in

mice do not necessarily lead to corresponding insights in humans — despite the many genetic similarities, humans are humans and mice are mice. But SNP analysis will without doubt prove revolutionary, and we can have every hope that the resulting discoveries about complex genetic traits will in time produce significant improvements in human health. ■

Joseph H. Nadeau is in the Department of Genetics, BRB 624, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106-4955, USA. e-mail: jhn4@po.cwru.edu

1. Moriawaki, K. in *Genetics in Wild Mice* (eds Moriawaki, K., Shiroishi, T. & Yonekawa, H.) xiii–xxv (Karger, New York, 1994).
2. Russell, E. S. in *Origins of Inbred Mice* (ed. Morse, H. C.) 33–43 (Academic, New York, 1978).
3. Nadeau, J. H. & Taylor, B. A. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
4. Makalowski, W. & Boguski, M. S. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
5. Mural, R. J. *et al. Science* **296**, 1661–1671 (2002).
6. Wade, C. M. *et al. Nature* **420**, 574–578 (2002).
7. Paigen, K. & Eppig, J. T. *Mammal. Genome* **11**, 715–717 (2000).
8. www.jax.org/phenome
9. Glazier, A. M., Nadeau, J. H. & Aitman, T. J. *Science* (in the press).
10. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
11. www.celeera.org
12. Weiss, K. M. & Clark, A. G. *Trends Genet.* **18**, 19–24 (2002).
13. Reich, D. E. *et al. Nature Genet.* **32**, 135–142 (2002).
14. Morse, H. C. (ed.) *Origins of Inbred Mice* (Academic, New York, 1978).
15. Ferris, S. D., Sage, R. D. & Wilson, A. C. *Nature* **295**, 163–165 (1982).
16. Bishop, C. E., Boursot, P., Baron, B., Bonhomme, F. & Hatat, D. *Nature* **325**, 70–72 (1985).
17. Lyon, M. F. *Annu. Rev. Genomics Hum. Genet.* **3**, 1–16 (2002).

most of the confirmed and 'predicted' human chromosome 21 genes. Reymond *et al.* then used three experimental approaches to detect their expression in mice, looking at 12 adult tissues and 6 developmental stages. This extensive survey forms the basis of the group's chromosome 21 'gene atlas'.

All three approaches used by Reymond *et al.* are based on detecting messenger RNA (because the first step in gene expression is the production of an mRNA copy of the gene). For 'in situ hybridization', a label is incorporated into a DNA or RNA molecule that is complementary to a given mRNA. The complementary molecule (cDNA or cRNA) is then incubated with a whole mouse embryo (for 'whole-mount in situ hybridization') or with a tissue section, and binds to the target mRNA. Developing the label produces a coloured product, revealing where the mRNA was present (Fig. 1). In an embryo, resolution is at the level of a tissue or organ. Tissue-section (histological) analysis provides higher resolution, to the level of cells, but with decreased sensitivity.

The third tactic used by Reymond *et al.* was to isolate mRNA from a tissue or organ; next, the mRNA was 'reverse transcribed' to produce a cDNA molecule, many copies of which were then generated by the polymerase chain reaction. This amplification technique, RT-PCR for short, is extremely sensitive, allowing the detection of even a few copies of an mRNA. Meanwhile, Gitton *et al.* used these same techniques to provide an independent look at one early developmental stage. They also extend Reymond *et al.*'s results by analysing the rapidly developing postnatal brain.

Several broad trends were evident in both data sets. For instance, during early embryonic development, many of the mouse orthologues of human chromosome 21 genes are expressed broadly, throughout many tissues; this pattern becomes more restricted as development proceeds. Reymond *et al.* also show that, in adult mice, the number of different genes expressed in a given tissue varies greatly: using RT-PCR, for example, they detect the expression of a surprising 85% of chromosome 21 genes in adult brain, but only 21% in muscle. Both groups also pinpoint specific genes that might be relevant to Down's syndrome, by virtue of their expression in the brain, limbs, heart or gut — structures commonly affected by triplication of chromosome 21.

Gitton *et al.* also looked further at the kinds of genes that are expressed broadly or in restricted patterns, by considering the mouse genes for which there are related sequences (paralogues) in a single-celled organism (yeast), in nematode worms and in fruitflies. They found that most of the mouse genes that had paralogues in all three lower organisms were expressed broadly or

Functional genomics

A time and place for every gene

Roger H. Reeves

One benefit of studying mice is that most of their genes have counterparts in humans. Two groups have used this similarity to study when and where the genes found on human chromosome 21 are switched on.

The publication of the complete DNA sequence of human chromosome 21 two years ago¹ provided the basis for identifying every one of the genes on this chromosome. On pages 582 and 586 of this issue, Reymond and colleagues² and Gitton and co-workers³ extend this tremendous accomplishment by analysing the expression of the mouse equivalents of these human genes in many different mouse tissues, and at several different stages of development. Chromosome 21 was chosen for these large-scale, functional-genomics analyses because of its importance in human development — Down's syndrome occurs when a person inherits three copies of chromosome 21 instead of the normal two.

Affecting one in 700 live births, Down's syndrome has a significant impact on afflicted people, their families and society. The disorder invariably results in cognitive

impairment and characteristic variations in digits and in the ridge formations on the hands and feet. Triplication of chromosome 21 is also a risk factor for congenital heart disease and Hirschsprung's disease, a colon disorder, as well as many other developmental abnormalities. But it remains unknown how inheriting three copies of chromosome 21 produces this broad spectrum of problems. Finding out when and where each chromosome 21 gene is expressed during development is a crucial step towards understanding the syndrome.

Of course, it's difficult to analyse gene expression during human development. But studies of mice are a valuable alternative, not least because one then has access to every tissue at all stages of development; moreover, many human genes have equivalents (orthologues) in mice. Reymond *et al.*² and Gitton *et al.*³ started by identifying orthologues of

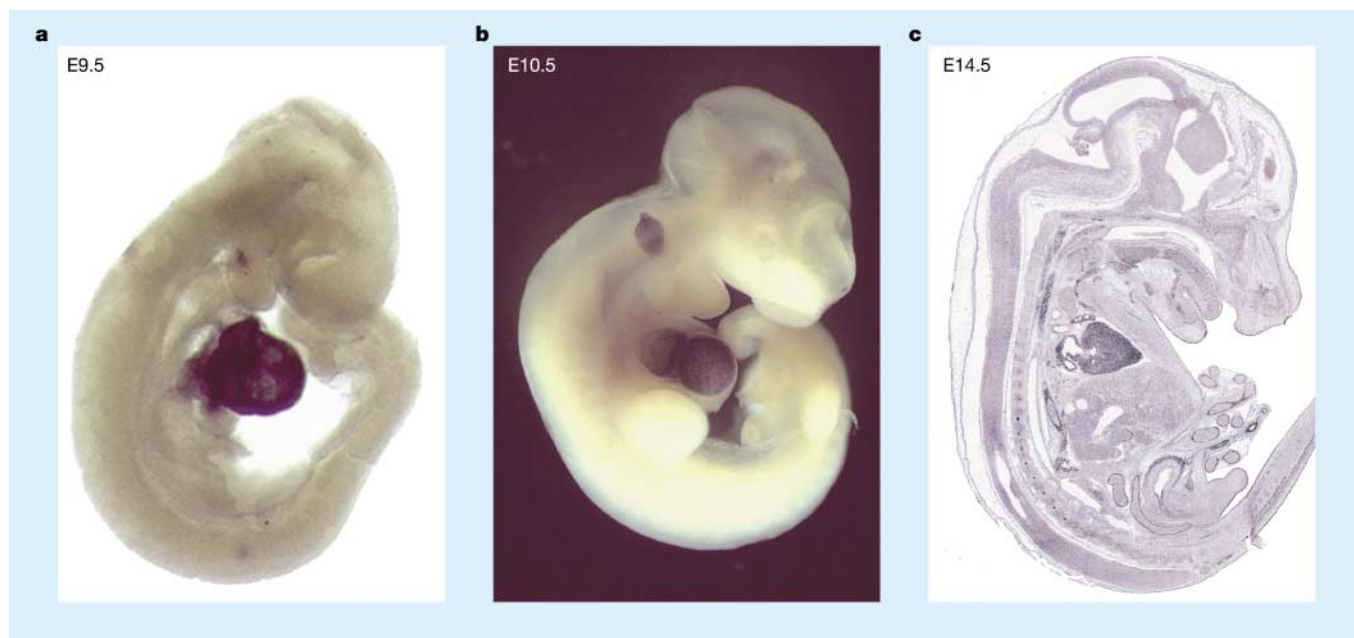


Figure 1 Analysing gene expression in the heart by *in situ* hybridization. a, b, Whole-mount *in situ* hybridization, showing expression of the mouse *Sh3bgr* gene, an orthologue of the human *SH3BGR* gene, at two different

stages of embryonic development. c, A tissue section of the mouse embryo, showing the expression of the *Sh3bgr* gene in developing heart. E, embryonic day of development. (a, b, provided by V. Marigo; c, provided by G. Eichele.)

ubiquitously in the mid-gestation embryo. Only a few showed an expression pattern that was restricted to specific tissues, suggesting that these genes might have general 'house-keeping' functions that are required in all cells. By contrast, many of the genes with paralogues in the two multicellular organisms but not in yeast were expressed in tissue-restricted patterns. As a group, these regionally patterned genes were involved in cell–cell communication and signal transduction (functions that are less elaborate in yeast).

In addition, Gitton *et al.* looked for functional clustering — the occurrence of groups of genes that are coordinately expressed in the same tissues; such genes might be involved in common processes. They used a computational method (described in ref. 4) to search for these functional gene clusters. To do this, the authors examined large existing gene-expression databases called expressed sequence tag (EST) libraries. ESTs are short nucleotide sequences, hundreds or thousands of which are obtained from randomly picked clones in a cDNA library; each cDNA (and therefore each EST) library represents the mRNAs expressed in a particular tissue at a specific point in time. Here, Gitton *et al.* identified 190 relevant EST libraries, in which the chromosome 21 genes were represented by 4,186 ESTs (on average, there were 22 ESTs per gene). Using a combination of correlation and matrix analyses, the authors identified eight functional clusters of genes that were expressed to a similar degree in different libraries, suggesting that the clustered genes are regulated together.

This approach is attractive in that it uses an objective statistical assessment to find far more complex patterns in large data sets than

can be observed by eye. But there are caveats. In theory, the procedure is minimally influenced by the size of the EST library, yet most of the eight gene clusters were seen only in the largest libraries. And, surprisingly, no clusters were found from the brain, despite the extensive representation of brain ESTs and the observation^{2,3} that many genes are coordinately expressed in restricted brain regions. This does not negate the value of the correlations, but does suggest that the method may be influenced by other, not immediately obvious factors.

Genome-sequence databases and EST libraries, coupled with appropriate search engines, have transformed the process of discovering genes. The development of centralized gene-expression databases and computational tools such as those described here is crucial for the more complex, multi-dimensional problem of tracking the expression of these genes through space and time. Several refinements will, however, be needed to allow such databases to capture all the information about gene expression. Users need a way of assessing the quality of the data, which is inevitably uneven in such a large resource and becomes more so if multiple investigators contribute data to a central compilation. For example, Reymond *et al.* found that 98–99% of the analysed genes showed detectable expression by whole-mount *in situ* hybridization after 9.5 days of embryonic development. But, using the same procedure, Gitton *et al.* found a signal for only 70% of genes at this stage. Standard controls, such as reference gene sets or relative statistics (for instance, the 98% versus 70% quoted here), will be needed to evaluate a given result.

It will also be important to improve on the static, two-dimensional images of whole embryos. In the laboratory, stained embryos must be rotated to allow them to be viewed from all angles. Software analogous to that used to present three-dimensional molecular structures, combined with three-dimensional imaging by optical sectioning or other approaches, may allow comparable manipulation of these images. On the computational side, the correlation approach to gene expression is but an early step along the road to obtaining methods that will relate patterns of gene expression to developmental events.

Nonetheless, gene-expression databases such as these^{2,3} are already valuable for analysing where and when genes come into play during development, and so will provide clues to what those genes do. Ultimately, these spatiotemporal patterns must be linked to structural and physiological changes that result from — and which themselves further alter — gene expression. The resulting knowledge of the roles of individual genes in specific developmental pathways must then be combined with an understanding of how development differs from normal in genetic disorders, before we can comprehend and, eventually, ameliorate these syndromes. ■

Roger H. Reeves is in the Department of Physiology and the McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205–2105, USA.

e-mail: rreeves@jhmi.edu

1. The Chromosome 21 Mapping and Sequencing Consortium *Nature* **405**, 311–319 (2000).
2. Reymond, A. *et al.* *Nature* **420**, 582–586 (2002).
3. Gitton, Y. *et al.* *Nature* **420**, 586–590 (2002).
4. Ewing, R. M. *et al.* *Genome Res.* **9**, 950–959 (1999).

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium*

*A list of authors and their affiliations appears at the end of the paper

The sequence of the mouse genome is a key informational tool for understanding the contents of the human genome and a key experimental tool for biomedical research. Here, we report the results of an international collaboration to produce a high-quality draft sequence of the mouse genome. We also present an initial comparative analysis of the mouse and human genomes, describing some of the insights that can be gleaned from the two sequences. We discuss topics including the analysis of the evolutionary forces shaping the size, structure and sequence of the genomes; the conservation of large-scale synteny across most of the genomes; the much lower extent of sequence orthology covering less than half of the genomes; the proportions of the genomes under selection; the number of protein-coding genes; the expansion of gene families related to reproduction and immunity; the evolution of proteins; and the identification of intraspecies polymorphism.

With the complete sequence of the human genome nearly in hand^{1,2}, the next challenge is to extract the extraordinary trove of information encoded within its roughly 3 billion nucleotides. This information includes the blueprints for all RNAs and proteins, the regulatory elements that ensure proper expression of all genes, the structural elements that govern chromosome function, and the records of our evolutionary history. Some of these features can be recognized easily in the human sequence, but many are subtle and difficult to discern. One of the most powerful general approaches for unlocking the secrets of the human genome is comparative genomics, and one of the most powerful starting points for comparison is the laboratory mouse, *Mus musculus*.

Metaphorically, comparative genomics allows one to read evolution's laboratory notebook. In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by nearly one substitution for every two nucleotides (see below) as well as by deletion and insertion. The divergence rate is low enough that one can still align orthologous sequences, but high enough so that one can recognize many functionally important elements by their greater degree of conservation. Studies of small genomic regions have demonstrated the power of such cross-species conservation to identify putative genes or regulatory elements^{3–12}. Genome-wide analysis of sequence conservation holds the prospect of systematically revealing such information for all genes. Genome-wide comparisons among organisms can also highlight key differences in the forces shaping their genomes, including differences in mutational and selective pressures^{13,14}.

Literally, comparative genomics allows one to link laboratory notebooks of clinical and basic researchers. With knowledge of both genomes, biomedical studies of human genes can be complemented by experimental manipulations of corresponding mouse genes to accelerate functional understanding. In this respect, the mouse is unsurpassed as a model system for probing mammalian biology and human disease^{15,16}. Its unique advantages include a century of genetic studies, scores of inbred strains, hundreds of spontaneous mutations, practical techniques for random mutagenesis, and, importantly, directed engineering of the genome through transgenic, knockout and knockin techniques^{17–22}.

For these and other reasons, the Human Genome Project (HGP) recognized from its outset that the sequencing of the human genome needed to be followed as rapidly as possible by the sequencing of the mouse genome. In early 2001, the International Human Genome Sequencing Consortium reported a draft sequence

covering about 90% of the euchromatic human genome, with about 35% in finished form¹. Since then, progress towards a complete human sequence has proceeded swiftly, with approximately 98% of the genome now available in draft form and about 95% in finished form.

Here, we report the results of an international collaboration involving centres in the United States and the United Kingdom to produce a high-quality draft sequence of the mouse genome and a broad scientific network to analyse the data. The draft sequence was generated by assembling about sevenfold sequence coverage from female mice of the C57BL/6J strain (referred to below as B6). The assembly contains about 96% of the sequence of the euchromatic genome (excluding chromosome Y) in sequence contigs linked together into large units, usually larger than 50 megabases (Mb).

With the availability of a draft sequence of the mouse genome, we have undertaken an initial comparative analysis to examine the similarities and differences between the human and mouse genomes. Some of the important points are listed below.

- The mouse genome is about 14% smaller than the human genome (2.5 Gb compared with 2.9 Gb). The difference probably reflects a higher rate of deletion in the mouse lineage.
- Over 90% of the mouse and human genomes can be partitioned into corresponding regions of conserved synteny, reflecting segments in which the gene order in the most recent common ancestor has been conserved in both species.
- At the nucleotide level, approximately 40% of the human genome can be aligned to the mouse genome. These sequences seem to represent most of the orthologous sequences that remain in both lineages from the common ancestor, with the rest likely to have been deleted in one or both genomes.
- The neutral substitution rate has been roughly half a nucleotide substitution per site since the divergence of the species, with about twice as many of these substitutions having occurred in the mouse compared with the human lineage.
- By comparing the extent of genome-wide sequence conservation to the neutral rate, the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be about 5%. This proportion is much higher than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions, regulatory elements, non-protein-coding genes, and chromosomal structural elements) under selection for biological function.
- The mammalian genome is evolving in a non-uniform manner,

with various measures of divergence showing substantial variation across the genome.

- The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new *de novo* gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.

- Dozens of local gene family expansions have occurred in the mouse lineage. Most of these seem to involve genes related to reproduction, immunity and olfaction, suggesting that these physiological systems have been the focus of extensive lineage-specific innovation in rodents.

- Mouse–human sequence comparisons allow an estimate of the rate of protein evolution in mammals. Certain classes of secreted proteins implicated in reproduction, host defence and immune response seem to be under positive selection, which drives rapid evolution.

- Despite marked differences in the activity of transposable elements between mouse and human, similar types of repeat sequences have accumulated in the corresponding genomic regions in both species. The correlation is stronger than can be explained simply by local (G+C) content and points to additional factors influencing how the genome is moulded by transposons.

- By additional sequencing in other mouse strains, we have identified about 80,000 single nucleotide polymorphisms (SNPs). The distribution of SNPs reveals that genetic variation among mouse strains occurs in large blocks, mostly reflecting contributions of the two subspecies *Mus musculus domesticus* and *Mus musculus musculus* to current laboratory strains.

The mouse genome sequence is freely available in public databases (GenBank accession number CAAA01000000) and is accessible through various genome browsers (http://www.ensembl.org/Mus_musculus/, <http://genome.ucsc.edu/> and <http://www.ncbi.nlm.nih.gov/genome/guide/mouse/>).

In this paper, we begin with information about the generation, assembly and evaluation of the draft genome sequence, the conservation of synteny between the mouse and human genomes, and the landscape of the mouse genome. We then explore the repeat sequences, genes and proteome of the mouse, emphasizing comparisons with the human. This is followed by evolutionary analysis of selection and mutation in the mouse and human lineages, as well as polymorphism among current mouse strains. A full and detailed description of the methods underlying these studies is provided as Supplementary Information. In many respects, the current paper is a companion to the recent paper on the human genome sequence¹. Extensive background information about many of the topics discussed below is provided there.

Background to the mouse genome sequencing project

Origins of the mouse

The precise origin of the mouse and human lineages has been the subject of recent debate. Palaeontological evidence has long indicated a great radiation of placental (eutherian) mammals about 65 million years ago (Myr) that filled the ecological space left by the extinction of the dinosaurs, and that gave rise to most of the eutherian orders²³. Molecular phylogenetic analyses indicate earlier divergence times of many of the mammalian clades. Some of these studies have suggested a very early date for the divergence of mouse from other mammals (100–130 Myr^{23–25}) but these estimates partially originate from the fast molecular clock in rodents (see below).

Recent molecular studies that are less sensitive to the differences in evolutionary rates have suggested that the eutherian mammalian radiation took place throughout the Late Cretaceous period (65–100 Myr), but that rodents and primates actually represent relatively late-branching lineages^{26,27}. In the analyses below, we use a divergence time for the human and mouse lineages of 75 Myr for the purpose of calculating evolutionary rates, although it is possible that the actual time may be as recent as 65 Myr.

Origins of mouse genetics

The origin of the mouse as the leading model system for biomedical research traces back to the start of human civilization, when mice became commensal with human settlements. Humans noticed spontaneously arising coat-colour mutants and recorded their observations for millennia (including ancient Chinese references to dominant-spotting, waltzing, albino and yellow mice). By the 1700s, mouse fanciers in Japan and China had domesticated many varieties as pets, and Europeans subsequently imported favourites and bred them to local mice (thereby creating progenitors of modern laboratory mice as hybrids among *M. m. domesticus*, *M. m. musculus* and other subspecies). In Victorian England, ‘fancy’ mice were prized and traded, and a National Mouse Club was founded in 1895 (refs 28, 29).

With the rediscovery of Mendel’s laws of inheritance in 1900, pioneers of the new science of genetics (such as Cuenot, Castle and Little) were quick to recognize that the discontinuous variation of fancy mice was analogous to that of Mendel’s peas, and they set out to test the new theories of inheritance in mice. Mating programmes were soon established to create inbred strains, resulting in many of the modern, well-known strains (including C57BL/6J)³⁰.

Genetic mapping in the mouse began with Haldane’s report³¹ in 1915 of linkage between the pink-eye dilution and albino loci on the linkage group that was eventually assigned to mouse chromosome 7, just 2 years after the first report of genetic linkage in *Drosophila*. The genetic map grew slowly over the next 50 years as new loci and linkage groups were added—chromosome 7 grew to three loci by 1935 and eight by 1954. The accumulation of serological and enzyme polymorphisms from the 1960s to the early 1980s began to fill out the genome, with the map of chromosome 7 harbouring 45 loci by 1982 (refs 29, 31).

The real explosion, however, came with the development of recombinant DNA technology and the advent of DNA-sequence-based polymorphisms. Initially, this involved the detection of restriction-fragment length polymorphisms (RFLPs)³²; later, the emphasis shifted to the use of simple sequence length polymorphisms (SSLPs; also called microsatellites), which could be assayed easily by polymerase chain reaction (PCR)^{33–36} and readily revealed polymorphisms between inbred laboratory strains.

Origins of mouse genomics

When the Human Genome Project (HGP) was launched in 1990, it included the mouse as one of its five central model organisms, and targeted the creation of genetic, physical and eventually sequence maps of the mouse genome.

By 1996, a dense genetic map with nearly 6,600 highly polymorphic SSLP markers ordered in a common cross had been developed³⁴, providing the standard tool for mouse genetics. Subsequent efforts filled out the map to over 12,000 polymorphic markers, although not all of these loci have been positioned precisely relative to one another. With these and other loci, Haldane’s original two-marker linkage group on chromosome 7 had now swelled to about 2,250 loci.

Physical maps of the mouse genome also proceeded apace, using sequence-tagged sites (STS) together with radiation-hybrid panels^{37,38} and yeast artificial chromosome (YAC) libraries to construct dense landmark maps³⁹. Together, the genetic and physical maps provide thousands of anchor points that can be used to tie

clones or DNA sequences to specific locations in the mouse genome.

Other resources included large collections of expressed-sequence tags (EST)⁴⁰, a growing number of full-length complementary DNAs^{41,42} and excellent bacterial artificial chromosome (BAC) libraries⁴³. The latter have been used for deriving large sets of BAC-end sequences³⁷ and, as part of this collaboration, to generate a fingerprint-based physical map⁴⁴. Furthermore, key mouse genome databases were developed at the Jackson (<http://www.informatics.jax.org/>), Harwell (<http://www.har.mrc.ac.uk/>) and RIKEN (<http://genome.rtc.riken.go.jp/>) laboratories to provide the community with access to this information.

With these resources, it became straightforward (but not always easy) to perform positional cloning of classic single-gene mutations for visible, behavioural, immunological and other phenotypes. Many of these mutations provide important models of human disease, sometimes recapitulating human phenotypes with uncanny accuracy. It also became possible for the first time to begin dissecting polygenic traits by genetic mapping of quantitative trait loci (QTL) for such traits.

Continuing advances fuelled a growing desire for a complete sequence of the mouse genome. The development of improved random mutagenesis protocols led to the establishment of large-scale screens to identify interesting new mutants, increasing the need for more rapid positional cloning strategies. QTL mapping experiments succeeded in localizing more than 1,000 loci affecting physiological traits, creating demand for efficient techniques capable of trawling through large genomic regions to find the underlying genes. Furthermore, the ability to perform directed mutagenesis of the mouse germ line through homologous recombination made it possible to manipulate any gene given its DNA sequence, placing an increasing premium on sequence information. In all of these cases, it was clear that genome sequence information could markedly accelerate progress.

Origin of the Mouse Genome Sequencing Consortium

With the sequencing of the human genome well underway by 1999, a concerted effort to sequence the entire mouse genome was organized by a Mouse Genome Sequencing Consortium (MGSC). The MGSC originally consisted of three large sequencing centres—the Whitehead/Massachusetts Institute of Technology (MIT) Center for Genome Research, the Washington University Genome Sequencing Center, and the Wellcome Trust Sanger Institute—together with an international database, Ensembl, a joint project between the European Bioinformatics Institute and the Sanger Institute.

In addition to the genome-wide efforts of the MGSC, other publicly funded groups have been contributing to the sequencing of the mouse genome in specific regions of biological interest. Together, the MGSC and these programmes have so far yielded clone-based draft sequence consisting of 1,859 Mb (74%, although there is redundancy) and finished sequence of 477 Mb (19%) of the mouse genome. Furthermore, Mural and colleagues⁴⁵ recently reported a draft sequence of mouse chromosome 16 containing 87 Mb (3.5%).

To analyse the data reported here, the MGSC was expanded to include the other publicly funded sequencing groups and a Mouse Genome Analysis Group consisting of scientists from 27 institutions in 6 countries.

Generating the draft genome sequence

Sequencing strategy

Sanger and co-workers developed the strategy of random shotgun sequencing in the early 1980s, and it has remained the mainstay of genome sequencing over the ensuing two decades. The approach involves producing random sequence 'reads', generating a preliminary assembly on the basis of sequence overlaps, and then perform-

ing directed sequencing to obtain a 'finished' sequence with gaps closed and ambiguities resolved⁴⁶. Ansorge and colleagues⁴⁷ extended the technique by the use of 'paired-end sequencing', in which sequencing is performed from both ends of a cloned insert to obtain linking information, which is then used in sequence assembly. More recently, Myers and co-workers⁴⁸, and others, have developed efficient algorithms for exploiting such linking information.

A principal issue in the sequencing of large, complex genomes has been whether to perform shotgun sequencing on the entire genome at once (whole-genome shotgun, WGS) or to first break the genome into overlapping large-insert clones and to perform shotgun sequencing on these intermediates (hierarchical shotgun)⁴⁶. The WGS technique has the advantage of simplicity and rapid early coverage; it readily works for simple genomes with few repeats, but there can be difficulties encountered with genomes that contain highly repetitive sequences (such as the human genome, which has near-perfect repeats spanning hundreds of kilobases). Hierarchical shotgun sequencing overcomes such difficulties by using local assembly, thus decreasing the number of repeat copies in each assembly and allowing comparison of large regions of overlaps between clones. Consequently, efforts to produce finished sequences of complex genomes have relied on either pure hierarchical shotgun sequencing (including those of *Caenorhabditis elegans*⁴⁹, *Arabidopsis thaliana*⁴⁹ and human¹) or a combination of WGS and hierarchical shotgun sequencing (including those of *Drosophila melanogaster*⁵⁰, human² and rice⁵¹).

The ultimate aim of the MGSC is to produce a finished, richly annotated sequence of the mouse genome to serve as a permanent reference for mammalian biology. In addition, we wished to produce a draft sequence as rapidly as possible to aid in the interpretation of the human genome sequence and to provide a useful intermediate resource to the research community. Accordingly, we adopted a hybrid strategy for sequencing the mouse genome. The strategy has four components: (1) production of a BAC-based physical map of the mouse genome by fingerprinting and sequencing the ends of clones of a BAC library⁴⁴; (2) WGS sequencing to approximately sevenfold coverage and assembly to generate an initial draft genome sequence; (3) hierarchical shotgun sequencing of BAC clones covering the mouse genome combined with the WGS data to create a hybrid WGS-BAC assembly; and (4) production of a finished sequence by using the BAC clones as a template for directed finishing. This mixed strategy was designed to exploit the simpler organizational aspects of WGS assemblies in the initial phase, while still culminating in the complete high-quality sequence afforded by clone-based maps.

We chose to sequence DNA from a single mouse strain, rather than from a mixture of strains⁴⁵, to generate a solid reference foundation, reasoning that polymorphic variation in other strains could be added subsequently (see below). After extensive consultation with the scientific community⁵², the B6 strain was selected because of its principal role in mouse genetics, including its well-characterized phenotype and role as the background strain on which many important mutations arose. We elected to sequence a female mouse to obtain equal coverage of chromosome X and autosomes. Chromosome Y was thus omitted, but this chromosome is highly repetitive (the human chromosome Y has multiple duplicated regions exceeding 100 kb in size with 99.9% sequence identity⁵³) and seemed an unwise target for the WGS approach. Instead, mouse chromosome Y is being sequenced by a purely clone-based (hierarchical shotgun) approach.

Sequencing and assembly

The genome assembly was based on a total of 41.4 million sequence reads derived from both ends of inserts (paired-end reads) of various clone types prepared from B6 female DNA. The inserts ranged in size from 2 to 200 kb (Table 1). The three large MGSC

Table 1 **Distribution of sequence reads**

Insert size (kb)*	Vector	Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
		All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
2	Plasmid	3.8	3.7	3.1	2.9	1.8	1.5	0.71	0.61	1.2
4	Plasmid	31.3	24.7	22.1	21.5	14.7	12.6	5.89	5.03	17.7
6	Plasmid	1.2	1.0	0.8	0.8	0.5	0.5	0.22	0.19	1.0
10	Plasmid	2.5	2.4	2.1	1.7	1.3	1.0	0.52	0.42	4.3
40	Fosmid	2.1	1.3	1.2	1.1	0.6	0.5	0.26	0.21	9.3
150–200	BAC	0.4	0.4	0.4	0.4	0.2	0.2	0.09	0.07	13.7
Other	Plasmid	0.07	0.05	0.03	0.04	0.03	0.03	0.01	0.01	0.02
Total		41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2
Centre		Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
		All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
Whitehead Institute		22.2	18.0	15.9	15.7	10.7	9.2	4.28	3.68	21.3
Washington University		11.5	8.3	7.5	7.1	4.7	3.9	1.87	1.57	5.9
Sanger Institute		6.7	6.3	5.4	4.7	3.3	2.7	1.31	1.09	5.3
University of Utah		0.6	0.6	0.5	0.5	0.3	0.3	0.13	0.11	1.0
The Institute for Genomic Research		0.5	0.4	0.4	0.4	0.2	0.2	0.09	0.08	13.7
Total		41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2

* The approximate mean size of inserts of various libraries. Each library was individually tracked and evaluated. Insert sizes were intended to cover a narrow range as determined empirically against assembled sequence.

† Bases refers to the bases present in the used reads after trimming for quality.

‡ Sequence coverage estimated on the basis of all used reads after trimming for quality and a 2.5-Gb euchromatic genome. This excludes the heterochromatic portion, which contains extensive arrays of tandemly repeated sequence such as that found in the centromeres, rDNA satellites and the *Sp100-rs* array.

§ Physical coverage refers to the total cloned DNA in the paired reads.

|| Consists of a small number of unpaired reads and BAC-based reads used for methods development and consistency checks.

sequencing centres generated 40.4 million reads, and 0.6 million reads were generated at the University of Utah. In addition, we used 0.4 million reads from both ends of BAC inserts reported by The Institute for Genome Research⁵⁴.

A total of 33.6 million reads passed extensive checks for quality and source, of which 29.7 million were paired; that is, derived from opposite ends of the same clone (Table 1). The assembled reads represent approximately 7.7-fold sequence coverage of the euchromatic mouse genome (6.5-fold coverage in bases with a Phred quality score of >20)⁵⁵. Together, the clone inserts provide roughly 47-fold physical coverage of the genome.

The sequence reads, together with the pairing information, were used as input for two recently developed sequence-assembly programs, Arachne^{56,57} and Phusion⁵⁸. No mapping information and no clone-based sequences were used in the WGS assembly, with the exception of a few reads (<0.1% of the total) derived from a handful of BACs, which were used as internal controls. The assembly programs were tested and compared on intermediate data sets over the course of the project and were thereby refined. The programs produced comparable outputs in the final assembly. The assembly generated by Arachne was chosen as the draft sequence described here because it yielded greater short-range and long-range continuity with comparable accuracy.

The assembly contains 224,713 sequence contigs, which are connected by at least two read-pair links into supercontigs (or scaffolds). There are a total of 7,418 supercontigs at least 2 kb in length, plus a further 37,125 smaller supercontigs representing

<1% of the assembly. The contigs have an N50 length of 24.8 kb, whereas the supercontigs have an N50 length that is approximately 700-fold larger at 16.9 Mb (N50 length is the size x such that 50% of the assembly is in units of length at least x). In fact, most of the genome lies in supercontigs that are extremely large: the 200 largest supercontigs span more than 98% of the assembled sequence, of which 3% is within sequence gaps (Table 2).

Anchoring to chromosomes

We assigned as many supercontigs as possible to chromosomal locations in the proper order and orientation. Supercontigs were localized largely by sequence alignments with the extensively validated mouse genetic map³⁴, with some additional localization provided by the mouse radiation-hybrid map³⁷ and the BAC map⁴⁴. We found no evidence of incorrect global joins within the supercontigs (that is, multiple markers supporting two discordant locations within the genome), and thus were able to place them directly. Altogether, we placed 377 supercontigs, including all supercontigs >500 kb in length.

Once much of the sequence was anchored, it was possible to exploit additional read-pair and physical mapping information to obtain greater continuity (Table 2). For example, some adjacent supercontigs were connected by BAC-end (or other) links, satisfying appropriate length and orientation constraints, including single links. Furthermore, some adjacent extended supercontigs were connected by means of fingerprint contigs in the BAC-based physical map. These additional links were used to join sequences

Table 2 **Basic statistics of the MGSCv3 assembly**

Features	Number	N50 length (kb)*	Bases (Gb)	Bases plus gaps (Gb)	Percentage of genome†
All anchored contigs‡	176,471	25.9	2.372	2.372	94.9
All anchored supercontigs	377	18,600	2.372	2.477	99.1
All ultracontigs	88	50,600	2.372	2.493	99.7
Unanchored contigs‡	48,242	2.3	0.106	0.106	–
Largest 200 supercontigs	200	18,700	2.352	2.455	98.2
Largest 100 supercontigs	100	22,900	1.955	2.039	81.6

* Not including gaps.

† Calculated on the basis of a 2.5-Gb euchromatic genome. Includes spanned gaps.

‡ The unanchored contigs, grouped into 44,166 unanchored supercontigs with an N50 value of 3.4 kb. The N50 value for all contigs is 24.8 kb, and for all supercontigs is 16,900 kb (excluding gaps). Inspection suggests that most of these unanchored contigs fall into gaps in the ultracontigs and are thus accounted for in the 'bases plus gaps' estimate.

into ultracontigs. In the end, a total of 88 ultracontigs with an N50 length of 50.6 Mb (exclusive of gaps) contained 95.7% of the assembled sequence (Fig. 1). Continuity near telomeres tends to be lower, and two chromosomes (5 and X) have unusually large numbers of ultracontigs.

Proportion of genome contained in the assembly

This was assessed by comparison with publicly available finished genome sequence and mouse cDNA sequences. Of the 187 Mb of finished mouse sequence, 96% was contained in the anchored assembly. This finished sequence, however, is not a completely random cross-section of the genome (it has been cloned as BACs, finished, and in some cases selected on the basis of its gene content). Of 11,452 cDNA sequences from the curated RefSeq collection, 99.3% of the cDNAs could be aligned to the genome sequence (see Supplementary Information). These alignments contained 96.4% of the cDNA bases. Together, this indicates that the draft genome sequence includes approximately 96% of the euchromatic portion of the mouse genome, with about 95% anchored (Table 1).

Genome size

On the basis of the estimated sizes of the ultracontigs and gaps between them, the total length of the euchromatic mouse genome

was estimated to be about 2.5 Gb (see Supplementary Information), or about 14% smaller than that of the euchromatic human genome (about 2.9 Gb) (Table 3). The ultracontigs include spanned gaps, whose lengths are estimated on the basis of paired-end reads and alignment against the human sequence (see below). To test the accuracy of the ultracontig lengths, we compared the actual length of 675 finished mouse BAC sequences (from the B6 strain) with the corresponding estimated length from the draft genome sequence. The ratio of estimated length to actual length had a median value of 0.9994, with 68% of cases falling within 0.99–1.01 and 84% of cases within 0.98–1.02.

Quality assessment at intermediate scale

Although no evidence of large-scale misassembly was found when anchoring the assembly onto the mouse chromosomes, we examined the assembly for smaller errors.

To assess the accuracy at an intermediate scale, we compared the positions of well-studied markers on the mouse genetic map and in the genome assembly (see Supplementary Information). Out of 2,605 genetic markers that were unambiguously mapped to the sequence assembly (BLAST match using 10^{-100} or better as an *E*-value to a single location) we found 1.8% in which the chromosomal assignment in the genetic map conflicted with that in the sequence. This is well within the known range of erroneous assignments within the genetic map³⁴. We tested 11 such discrepant markers by re-mapping them in a mouse cross. In ten cases, the data showed that the previous genetic map assignment was erroneous and supported the position in the draft sequence. In one case, the data supported the previous genetic map assignment and contradicted the assembly. By studying the one erroneous case, we recognized that a single 36-kb segment had been erroneously merged into a sequence contig by means of a single overlap of two reads. We screened the entire assembly for similar instances, affecting regions of at least 20 kb. Only 17 additional cases were found, with a median size of the incorrectly merged segment of 34 kb. These are being corrected in the next release of the MGSC sequence. We are continuing to investigate instances involving smaller incorrectly merged segments.

We also found 19 instances (0.7%) of conflicts in local marker order between the genetic map and sequence assembly. A conflict was defined as any instance that would require changing more than a single genotype in the data underlying the genetic map to resolve. We studied ten cases by re-mapping the genetic markers, and eight were found to be due to errors in the genetic map. On the basis of this analysis, we estimate that chromosomal misassignment and local misordering affects <0.3% of the assembled sequence.

Quality assessment at fine scale

We also assessed fine-scale accuracy of the assembly by carefully aligning it to about 10 Mb of finished BAC-derived sequence from the B6 strain. This revealed a total of 39 discrepancies of ≥ 50 bp in length (median size of 320 bp), reflecting small misassemblies either in the draft sequence or the finished BAC sequences. These discrepancies typically occurred at the ends of contigs in the WGS assembly, indicating that they may represent the incorrect incorporation of a single terminal read.

At the single nucleotide level in the assembly, the observed discrepancy rates varied in a manner consistent with the quality scores assigned to the bases in the WGS assembly (see Supplementary Information). Overall, 96% of nucleotides in the assembly have Arachne quality scores ≥ 40 , corresponding to a predicted error rate of 1 per 10,000 bases. Such bases had an observed discrepancy rate against finished sequence of 0.005%, or 5 errors per 100,000 bases.

Comparison with the draft sequence of chromosome 16

We also compared the sequence reported here to a draft sequence of mouse chromosome 16 recently published by Mural and

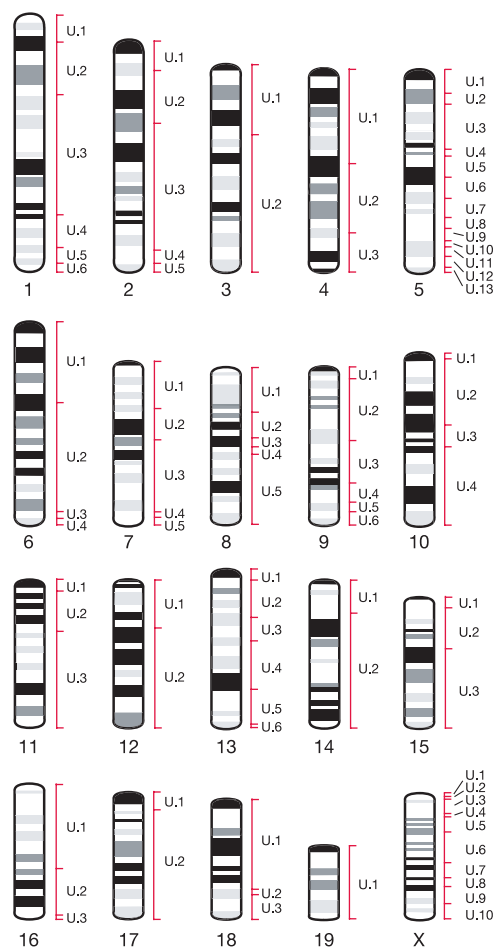


Figure 1 The mouse genome in 88 sequence-based ultracontigs. The position and extent of the 88 ultracontigs of the MGSCv3 assembly are shown adjacent to ideograms of the mouse chromosomes. All mouse chromosomes are acrocentric, with the centromeric end at the top of each chromosome. The supercontigs of the sequence assembly were anchored to the mouse chromosomes using the MIT genetic map. Neighbouring supercontigs were linked together into ultracontigs using information from single BAC links and the fingerprint and radiation-hybrid maps, resulting in 88 ultracontigs containing 95% of the bases in the euchromatic genome.

co-workers⁴⁵. Because the latter was produced from strain 129 and other mouse strains, it is expected to differ slightly at the nucleotide level but should otherwise show good agreement. The sequences align well at large scales (hundreds of kilobases), although the assembly by Mural and co-workers contains less total sequence (87 compared with 91 Mb) and includes a region of approximately 300 kb that we place on chromosome X. There were differences at intermediate scales, with our draft sequence showing better agreement with finished BAC-derived sequences (approximately fourfold fewer discrepancies of length ≥ 500 bp; 20 compared with 5 in about 2.8 Mb of finished sequence). These could not be explained by strain differences, as similar results were seen with finished sequence from the B6 and 129 strains.

Collapse of duplicated regions

The human genome contains many large duplicated regions, estimated to comprise roughly 5% of the genome⁵⁹, with nearly identical sequence. If such regions are also common in the mouse genome, they might collapse into a single copy in the WGS assembly. Such artefactual collapse could be detected as regions with unusually high read coverage, compared with the average depth of 7.4-fold in long assembled contigs. We searched for contigs that were >20 kb in size and contained >10 kb of sequence in which the read coverage was at least twofold higher than the average. Such regions comprised only a tiny fraction (<0.0001) of the total assembly, of which only half had been anchored to a chromosome. None of these windows had coverage exceeding the average by more than threefold. This may indicate that the mouse genome contains fewer large regions of near-exact duplication than the human. Alternatively, regions of near-exact duplication may have been systematically excluded by the WGS assembly programme. This issue is better addressed through hierarchical shotgun than WGS sequencing and will be examined more carefully in the course of producing a finished mouse genome sequence.

Unplaced reads and large tandem repeats

We expected that highly repetitive regions of the genome would not be assembled or would not be anchored on the chromosomes.

Indeed, 5.9 million of the 33.6 million passing reads were not part of anchored sequence, with 88% of these not assembled into sequence contigs and 12% assembled into small contigs but not chromosomally localized.

A striking example of unassembled sequence is a large region on mouse chromosome 1 that contains a tandem expansion of sequence containing the *Sp100-rs* gene fusion. This region is highly variable among mouse species and even laboratory strains, with estimated lengths ranging from 6 to 200 Mb^{60,61}. The bulk of this region was not reliably assembled in the draft genome sequence. The individual sequence reads together were found to contain 493-fold coverage of the *Sp100-rs* gene, suggesting that there are roughly 60 copies in the B6 genome (corresponding to a region of about 6 Mb). This is consistent with an estimate of 50 copies in B6 obtained by Southern blotting⁶².

We also examined centromeric sequences, including the euchromatin-proximal major satellite repeat (234 bases) and the telomere-proximal minor repeat (120 bases) found on some chromosomes^{63,64}. (Note that mouse chromosomes are all acrocentric, meaning that the centromere is adjacent to one telomere.) The minor satellite was poorly represented among the sequence reads (present in about 24,000 reads or $<0.1\%$ of the total) suggesting that this satellite sequence is difficult to isolate in the cloning systems used. The major satellite was found in about 3.6% of the reads; this is also lower than previous estimates based on density gradient experiments, which found that major satellites comprise about 5.5% of the mouse genome, or approximately 8 Mb per chromosome⁶⁵.

Evaluation of WGS assembly strategy

The WGS assembly described here involved only random reads, without any additional map-based information. By many criteria, the assembly is of very high quality. The N50 supercontig size of 16.9 Mb far exceeds that achieved by any previous WGS assembly, and the agreement with genome-wide maps is excellent. The assembly quality may be due to several factors, including the use of high-quality libraries, the variety of insert lengths in multiple

Table 3 Mouse chromosome size estimates

Chromosome	Actual bases in sequence (Mb)	Ultracontigs (Mb)		Gaps within supercontigs		Gaps between supercontigs					Total estimated size (Mb)‡
		Number	N50 size	Number	Mb	Captured by additional read pairs		Captured by fingerprint contigs*		Uncaptured†	
						Number	Mb	Number	Mb	Number	
All	2,372	88	52.7	176,094	104.5	252	14.0	37	2.30	68	2,493
1	183	6	52.7	13,178	7.8	16	1.1	1	0.32	5	192
2	169	5	111.1	12,141	6.5	4	0.1	1	0.20	4	176
3	149	2	108.9	10,630	6.8	17	0.7	3	0.16	1	157
4	140	3	83.1	10,745	6.3	14	0.4	3	0.26	2	147
5	137	13	17.8	11,288	6.7	11	0.5	3	0.11	12	144
6	138	4	91.4	10,021	6.6	19	1.1	2	0.26	3	146
7	122	5	45.1	9,484	5.7	55	3.4	4	0.12	4	131
8	119	5	35.0	9,186	6.1	7	0.2	2	0.12	4	125
9	116	6	26.8	8,479	4.5	6	0.6	1	0.06	5	121
10	121	4	50.4	9,490	5.4	9	0.6	0	0	3	127
11	115	3	80.4	8,681	4.3	2	0.0	1	0.05	2	119
12	105	2	77.4	7,577	4.0	27	1.2	2	0.00	1	110
13	107	6	28.0	7,910	4.7	13	0.8	4	0.19	5	113
14	107	2	93.6	7,605	4.0	10	0.5	2	0.12	1	112
15	96	3	65.3	7,025	4.3	2	0.1	0	0	2	100
16	91	3	62.3	6,695	4.4	1	0.0	0	0	2	95
17	85	2	80.8	6,584	3.7	17	1.2	4	0.19	1	90
18	84	3	73.5	6,192	3.2	2	0.0	0	0	2	87
19	55	1	57.7	3,934	2.4	7	0.6	2	0.12	0	58
X	134	10	19.9	9,249	7.0	13	0.8	2	0.00	9	142

* These gaps had fingerprint contigs spanning them. The size for 18 out of 37 were estimated using conserved synteny to determine the size of the region in the human genome. The remaining gaps were arbitrarily given the average size of the assessed gaps (59 kb), adjusted to reflect the 16% difference in genome size.

† Uncaptured gaps were estimated by mouse-human synteny to have a total size of 5 Mb. However, because some of these gaps are due to repetitive expansions in mouse (absent in human), the actual total for the uncaptured gaps is probably substantially higher. For example, one large uncaptured gap on chromosome 1 (the *Sp-100rs* region) is roughly 6 Mb (see text).

‡ Omitting centromeres and telomeres. These would add, on average, approximately 8 Mb per chromosome, or about 160 Mb to the genome. Also omitting uncaptured gaps between supercontigs.

libraries, the improved assembly algorithms, and the inbred nature of the mouse strain (in contrast to the polymorphisms in the human genome sequences). Another contributing factor may be that the mouse differs from the human in having less recent segmental duplication to confound assembly.

Notwithstanding the high quality of the draft genome sequence, we are mindful that it contains many gaps, small misassemblies and nucleotide errors. It is likely that these could not all be resolved by further WGS sequencing, therefore directed sequencing will be needed to produce a finished sequence. The results also suggest that WGS sequencing may suffice for large genomes for which only draft sequence is required, provided that they contain minimal amounts of sequence associated with recent segmental duplications or large, recent interspersed repeat elements.

Adding finished sequence

As a final step, we enhanced the WGS sequence assembly by substituting available finished BAC-derived sequence from the B6 strain. In total, we replaced 3,528 draft sequence contigs with 48.2 Mb of finished sequence from 210 finished BACs available at the time of the assembly. The resulting draft genome sequence, MGSCv3, was submitted to the public databases and is freely available in electronic form through various sources (see below).

The sequence data and assemblies have been freely available throughout the course of the project. The next step of the project, which is already underway, is to convert the draft sequence into a finished sequence. As the MGSC produces additional BAC assemblies and finished sequence, we plan to continue to revise and release enhanced versions of the genome sequence *en route* to a completely finished sequence⁶⁶, thereby providing a permanent foundation for biomedical research in the twenty-first century.

Conservation of synteny between mouse and human genomes

With the draft sequence in hand, we began our analysis by investigating the strong conservation of synteny between the mouse and human genomes. Beyond providing insight into evolutionary events that have moulded the chromosomes, this analysis facilitates further comparisons between the genomes.

Starting from a common ancestral genome approximately 75 Myr, the mouse and human genomes have each been shuffled by chromosomal rearrangements. The rate of these changes, however, is low enough that local gene order remains largely intact. It is thus possible to recognize syntenic (literally 'same thread') regions in the two species that have descended relatively intact from the common ancestor.

The earliest indication that genes reside in similar relative positions in different mammalian species traces to the observation that the albino and pink-eye dilution mutants are genetically closely linked in both mouse and rat^{67,68}. Significant experimental evidence came from genetic studies of somatic cells⁶⁹. In 1984, Nadeau and Taylor⁷⁰ used mouse linkage data and human cytogenetic data to compare the chromosomal locations of orthologous genes. On the

basis of a small data set (83 loci), they extrapolated that the mouse and human genomes could be parsed into roughly 180 syntenic regions. During two decades of subsequent work, the density of the synteny map has been increased, but the estimated number of syntenic regions has remained close to the original projection. A recent gene-based synteny map³⁷ used more than 3,600 orthologous loci to define about 200 regions of conserved synteny. However, it is recognized that such maps might still miss regions owing to insufficient marker density.

With a robust draft sequence of the mouse genome and >90% finished sequence of the human genome in hand, it is possible to undertake a more comprehensive analysis of conserved synteny. Rather than simply relying on known human–mouse gene pairs, we identified a much larger set of orthologous landmarks as follows. We performed sequence comparisons of the entire mouse and human genome sequences using the PatternHunter program⁷¹ to identify regions having a similarity score exceeding a high threshold (>40, corresponding to a minimum of a 40-base perfect match, with penalties for mismatches and gaps), with the additional property that each sequence is the other's unique match above this threshold. Such regions probably reflect orthologous sequence pairs, derived from the same ancestral sequence.

About 558,000 orthologous landmarks were identified; in the mouse assembly, these sequences have a mean spacing of about 4.4 kb and an N50 length of about 500 bp. The landmarks had a total length of roughly 188 Mb, comprising about 7.5% of the mouse genome. It should be emphasized that the landmarks represent only a small subset of the sequences, consisting of those that can be aligned with the highest similarity between the mouse and human genomes. (Indeed, below we show that about 40% of the human genome can be aligned confidently with the mouse genome.)

The locations of the landmarks in the two genomes were then compared to identify regions of conserved synteny. We define a syntenic segment to be a maximal region in which a series of landmarks occur in the same order on a single chromosome in both species. A syntenic block in turn is one or more syntenic segments that are all adjacent on the same chromosome in human and on the same chromosome in mouse, but which may otherwise be shuffled with respect to order and orientation. To avoid small artefactual syntenic segments owing to imperfections in the two draft genome sequences, we only considered regions above 300 kb and ignored occasional isolated interruptions in conserved order (see Supplementary Information). Thus, some small syntenic segments have probably been omitted—this issue will be addressed best when finished sequences of the two genomes are completed.

Marked conservation of landmark order was found across most of the two genomes (Fig. 2). Each genome could be parsed into a total of 342 conserved syntenic segments. On average, each landmark resides in a segment containing 1,600 other landmarks. The segments vary greatly in length, from 303 kb to 64.9 Mb, with a mean of 6.9 Mb and an N50 length of 16.1 Mb. In total, about 90.2% of the human genome and 93.3% of the mouse genome

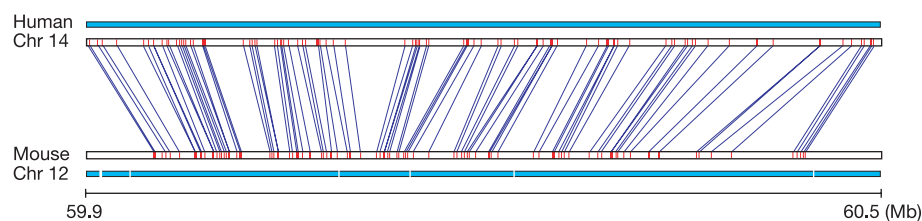


Figure 2 Conservation of synteny between human and mouse. We detected 558,000 highly conserved, reciprocally unique landmarks within the mouse and human genomes, which can be joined into conserved syntenic segments and blocks (defined in text). A typical 510-kb segment of mouse chromosome 12 that shares common ancestry with a

600-kb section of human chromosome 14 is shown. Blue lines connect the reciprocal unique matches in the two genomes. The cyan bars represent sequence coverage in each of the two genomes for the regions. In general, the landmarks in the mouse genome are more closely spaced, reflecting the 14% smaller overall genome size.

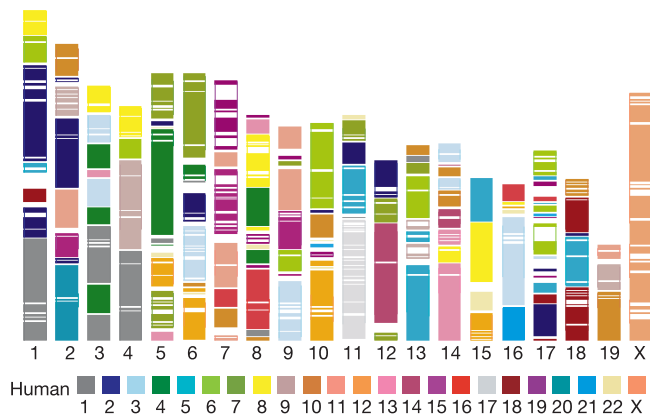


Figure 3 Segments and blocks >300 kb in size with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome. The 342 segments are separated from each other by thin, white lines within the 217 blocks of consistent colour.

unambiguously reside within conserved syntenic segments. The segments can be aggregated into a total of 217 conserved syntenic blocks, with an N50 length of 23.2 Mb.

The nature and extent of conservation of synteny differs substantially among chromosomes (Fig. 3 and Table 4). In accordance with expectation, the X chromosomes are represented as single, reciprocal syntenic blocks⁷². Human chromosome 20 corresponds entirely to a portion of mouse chromosome 2, with nearly perfect conservation of order along almost the entire length, disrupted only by a small central segment (Fig. 4a, d). Human chromosome 17 corresponds entirely to a portion of mouse chromosome 11, but extensive rearrangements have divided it into at least 16 segments (Fig. 4b, e). Other chromosomes, however, show evidence of much more extensive interchromosomal rearrangement than these cases (Fig. 4c, f).

We compared the new sequence-based map of conserved synteny with the most recent previous map based on 3,600 loci³⁰. The new map reveals many more conserved syntenic segments (342 com-

pared with 202) but only slightly more conserved syntenic blocks (217 compared with 170). Most of the conserved syntenic blocks had previously been recognized and are consistent with the new map, but many rearrangements of segments within blocks had been missed (notably on the X chromosome).

The occurrence of many local rearrangements is not surprising. Compared with interchromosomal rearrangements (for example, translocations), paracentric inversions (that is, those within a single chromosome and not including the centromere) carry a lower selective disadvantage in terms of the frequency of aneuploidy among offspring. These are also seen at a higher frequency in genera such as *Drosophila*, in which extensive cytogenetic comparisons have been carried out^{73,74}.

The block and segment sizes are broadly consistent with the random breakage model of genome evolution⁷⁵ (Fig. 5). At this gross level, there is no evidence of extensive selection for gene order across the genome. Selection in specific regions, however, is by no means excluded, and indeed seems probable (for example, for the major histocompatibility complex). Moreover, the analysis does not exclude the possibility that chromosomal breaks may tend to occur with higher frequency in some locations.

With a map of conserved syntenic segments between the human and mouse genomes, it is possible to calculate the minimal number of rearrangements needed to 'transform' one genome into the other^{70,76,77}. When applied to the 342 syntenic segments above, the most parsimonious path has 295 rearrangements. The analysis suggests that chromosomal breaks may have a tendency to reoccur in certain regions. With only two species, however, it is not yet possible to recover the ancestral chromosomal order or reconstruct the precise pathway of rearrangements. As more mammalian species are sequenced, it should be possible to draw such inferences and study the nature of chromosome rearrangement.

Genome landscape

We next sought to analyse the contents of the mouse genome, both in its own right and in comparison with corresponding regions of the human genome. The poster included with this issue provides a high-level view of the mouse genome, showing such features as genes and gene predictions, repetitive sequence content, (G+C) content, synteny with the human genome, and mouse QTLs.

Table 4 Syntenic properties of human and mouse chromosomes

Chromosome	Human			Mouse			Size (mouse/human)*
	Blocks	Segments	Fraction of chromosome in segments	Blocks	Segments	Fraction of chromosome in segments	
1	11	19	0.87	14	21	0.93	0.90
2	18	28	0.93	10	21	0.96	0.88
3	16	27	0.92	10	15	0.97	0.92
4	9	11	0.97	9	13	0.99	0.95
5	18	19	0.97	16	24	0.93	0.83
6	11	18	0.94	17	23	0.91	0.93
7	20	26	0.87	11	23	0.82	0.93
8	16	19	0.90	15	21	0.94	0.89
9	11	17	0.82	10	17	0.93	0.86
10	13	18	0.90	9	16	0.95	0.92
11	9	10	0.93	10	27	0.94	0.89
12	7	17	0.94	8	10	0.96	0.92
13	9	9	0.96	12	14	0.92	0.90
14	5	5	0.98	18	18	0.96	0.89
15	5	17	0.87	4	8	0.96	0.88
16	4	6	0.89	7	9	0.96	0.90
17	1	16	0.85	17	20	0.80	0.85
18	10	12	0.87	14	19	0.96	0.92
19	10	17	0.55	5	7	0.93	0.89
20	1	3	0.93	NA	NA	NA	NA
21	3	3	0.87	NA	NA	NA	NA
22	9	9	0.84	NA	NA	NA	NA
X	1	16	0.87	1	16	0.92	1.03
Total	217	342	0.90	217	342	0.93	0.91

NA, not applicable, as mouse has only 19 autosomes.

*Mean size ratio (mouse/human) on the basis of orthologous 100-kb mouse windows.

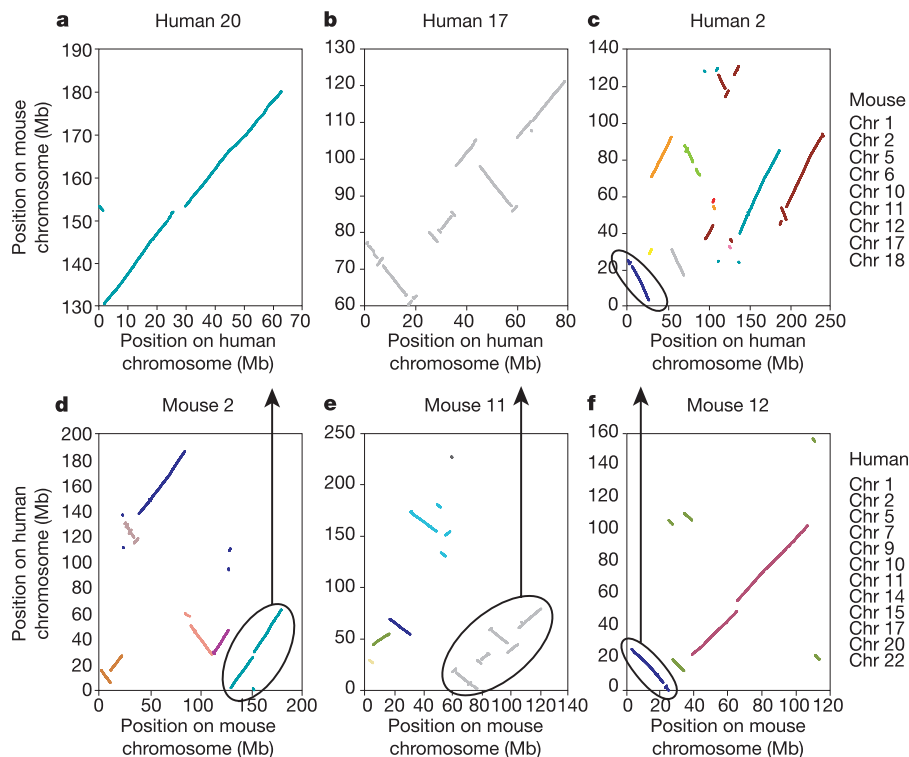


Figure 4 Dot plots of conserved syntenic segments in three human and three mouse chromosomes. For each of three human (a–c) and mouse (d–f) chromosomes, the positions of orthologous landmarks are plotted along the x axis and the corresponding position of the landmark on chromosomes in the other genome is plotted on the y axis. Different chromosomes in the corresponding genome are differentiated with distinct colours. In a remarkable example of conserved synteny, human chromosome 20 (a)

consists of just three segments from mouse chromosome 2 (d), with only one small segment altered in order. Human chromosome 17 (b) also shares segments with only one mouse chromosome (11) (e), but the 16 segments are extensively rearranged. However, most of the mouse and human chromosomes consist of multiple segments from multiple chromosomes, as shown for human chromosome 2 (c) and mouse chromosome 12 (f). Circled areas and arrows denote matching segments in mouse and human.

All of the mouse genome information is accessible in electronic form through various browsers: Ensembl (<http://www.ensembl.org>), the University of California at Santa Cruz (<http://genome.ucsc.edu>) and the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). These browsers allow users to scroll along the chromosomes and zoom in or out to any scale, as well as to display information at any desired level of detail. The mouse genome information has also been integrated into existing human genome browsers at these same organizations. In this section, we compare general properties of the mouse and human genomes.

Genome expansion and contraction

The projected total length of the euchromatic portion of the mouse genome (2.5 Gb) is about 14% smaller than that of the human genome (2.9 Gb). To investigate the source of this difference, we examined the relative size of intervals between consecutive orthologous landmarks in the human and mouse genomes. The mouse/human ratio has a mean at 0.91 for autosomes, but varies widely, with the mouse interval being larger than the human in 38% of cases (Fig. 6). Chromosome X, by contrast, shows no net relative expansion or contraction, with a mouse/human ratio of 1.03 (Fig. 6 and Table 4). What accounts for the smaller size of the mouse genome? We address this question below in the sections on repeat sequences and on genome evolution.

(G+C) content

The overall distribution of local (G+C) content is significantly different between the mouse and human genomes (Fig. 7). Such differences have been noted in biochemical studies^{78–81} and in comparative analyses of fourfold degenerate sites in codons of

mouse and human genes^{82–85}, but the availability of nearly complete genome sequences provides the first detailed picture of the phenomenon.

The mouse has a slightly higher overall (G+C) content than the human (42% compared with 41%), but the distribution is tighter. When local (G+C) content is measured in 20-kb windows across the genome, the human genome has about 1.4% of the windows with (G+C) content >56% and 1.3% with (G+C) content <33%. Such extreme deviations are virtually absent in the mouse genome. The contrast is even seen at the level of entire chromosomes. The

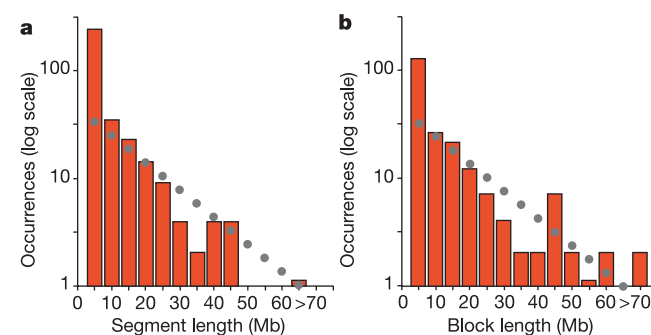


Figure 5 Size distribution of segments and blocks with syntenic conservation between mouse and human. **a**, **b**, The number of segments (**a**) and blocks (**b**) with syntenic conservation between mouse and human in 5-Mb bins (starting with 0.3–5 Mb) is plotted on a logarithmic scale. The dots indicate the expected values for the exponential curve of random breakage given the number of blocks and segments, respectively.

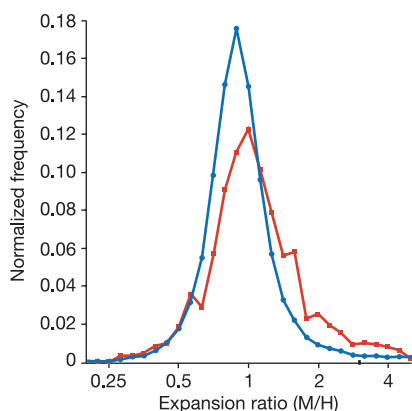


Figure 6 Size ratio of mouse to human for orthologous 100-kb windows. For each 100-kb region of the mouse genome, the size ratio to the related segment of the human genome was determined. The frequency of the various ratios is plotted on a logarithmic scale for both the autosomes (blue line) and the X chromosome (red line). The ratio for autosomes shows a mean of 0.91 but the ratio varies widely, with the mouse genome larger for 38% of the intervals. The X chromosome by contrast has a mean ratio of just over 1.0. Indeed, chromosome X is slightly smaller in human.

human has extreme outliers with respect to (G+C) content (the most extreme being chromosome 19), whereas the mouse chromosomes tend to be far more uniform (Fig. 8).

There is a strong positive correlation in local (G+C) content between orthologous regions in the mouse and human genomes (Fig. 9), but with the mouse regions showing a clear tendency to be less extreme in (G+C) content than the human regions. This tendency is not uniform, with the most extreme differences seen at the tails of the distribution.

In mammalian genomes, there is a positive correlation between gene density and (G+C) content^{81,86–89}. Given the differences in (G+C) content between human and mouse, we compared the distribution of genes—using the sets of orthologous mouse and human genes described below—with respect to (G+C) content for both genomes (Fig. 9). The density of genes differed markedly when expressed in terms of absolute (G+C) content, but was nearly

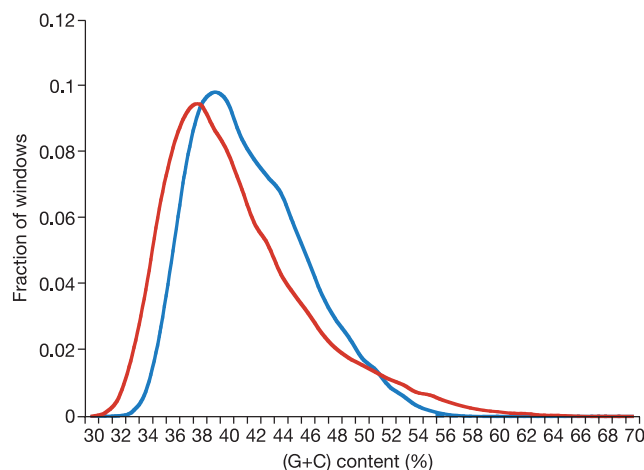


Figure 7 Distribution of (G+C) content in the mouse (blue) and human (red) genomes. Mouse has a higher mean (G+C) content than human (42% compared with 41%), but human has a larger fraction of windows with either high or low (G+C) content. The distribution was determined using the unmasked genomes in 20-kb non-overlapping windows, with the fraction of windows (y axis) in each percentage bin (x axis) plotted for both human and mouse.

identical when expressed in terms of percentiles of (G+C) content (Fig. 9). For example, both species have 75–80% of genes residing in the (G+C)-richest half of their genome. Mouse and human thus show similar degrees of homogeneity in the distribution of genes, despite the overall differences in (G+C) content. Notably, the mouse shows similar extremes of gene density despite being less extreme in (G+C) content.

What accounts for the differences in (G+C) content between mouse and human? Does it reflect altered selection for (G+C) content^{90,91}, altered mutational or repair processes^{92–94}, or possibly both? Data from additional species will probably be needed to address these issues. Any explanation will need to account for various mysterious phenomena. For example, although overall (G+C) content in mouse is slightly higher than in human (42% compared with 41%), the (G+C) content of chromosome X is slightly lower (39.0% compared with 39.4%). The effect is even more pronounced if one excludes lineage-specific repeats (see below), thereby focusing primarily on shared DNA. In that case, mouse autosomes have an overall (G+C) content that is 1.5% higher than human autosomes (41.2% compared with 39.7%) whereas mouse chromosome X has a (G+C) content that is 1% lower than human chromosome X (37.8% compared with 36.8%).

CpG islands

In mammalian genomes, the palindromic dinucleotide CpG is usually methylated on the cytosine residue. Methyl-CpG is mutated by deamination to TpG, leading to approximately fivefold under-representation of CpG across the human^{1,95} and mouse genomes. In some regions of the genome that have been implicated in gene regulation, CpG dinucleotides are not methylated and thus are not subject to deamination and mutation. Such regions, termed CpG islands, are usually a few hundred nucleotides in length, have high (G+C) content and above average representation of CpG dinucleotides.

We applied a computer program that attempts to recognize CpG islands on the basis of (G+C) and CpG content of arbitrary lengths of sequence^{96,97} to the non-repetitive portions of human and mouse genome sequences (see Supplementary Information). The mouse genome contains fewer CpG islands than the human genome (about 15,500 compared with 27,000), which is qualitatively consistent with previous reports⁹⁸. The absolute number of islands identified

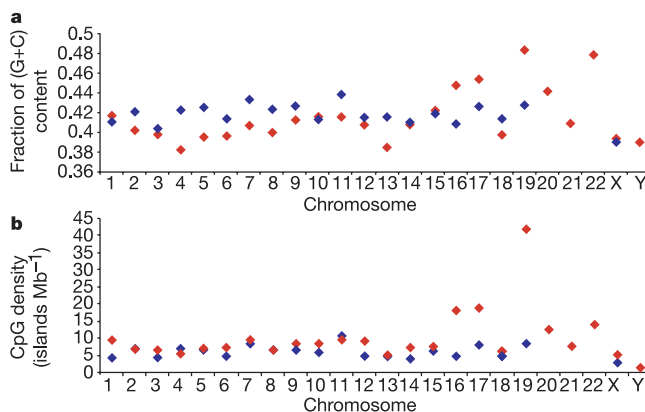


Figure 8 (G+C) content and density of CpG islands shows more variability in human (red) than mouse (blue) chromosomes. **a**, The (G+C) content for each of the mouse chromosomes is relatively similar, whereas human chromosomes show more variation; chromosomes 16, 17, 19 and 22 have higher (G+C) content, and chromosome 13 lower (G+C) content. **b**, Similarly, the density of CpG islands is relatively homogenous for all mouse chromosomes and more variable in human, with the same exceptions. Note that the mouse and human chromosomes are matched by chromosome number, not by regions of conserved synteny.

depends on the precise definition of a CpG island used, but the ratio between the two species remains fairly constant.

The reason for the smaller number of predicted CpG islands in mouse may relate simply to the smaller fraction of the genome with extremely high (G+C) content⁹⁹ and its effect on the computer algorithm. Approximately 10,000 of the predicted CpG islands in each species show significant sequence conservation with CpG islands in the orthologous intervals in the other species, falling within the orthologous landmarks described above. Perhaps these represent functional CpG islands, a proposition that can now be tested experimentally⁸⁴.

Repeats

The single most prevalent feature of mammalian genomes is their repetitive sequences, most of which are interspersed repeats representing ‘fossils’ of transposable elements. Transposable elements are a principal force in reshaping the genome, and their fossils thus provide powerful reporters for measuring evolutionary forces acting on the genome. A recent paper on the human genome sequence¹ provided extensive background on mammalian transposons, describing their biology and illustrating many applications to evolutionary studies. Here, we will focus primarily on comparisons between the repeat content of the mouse and human genomes.

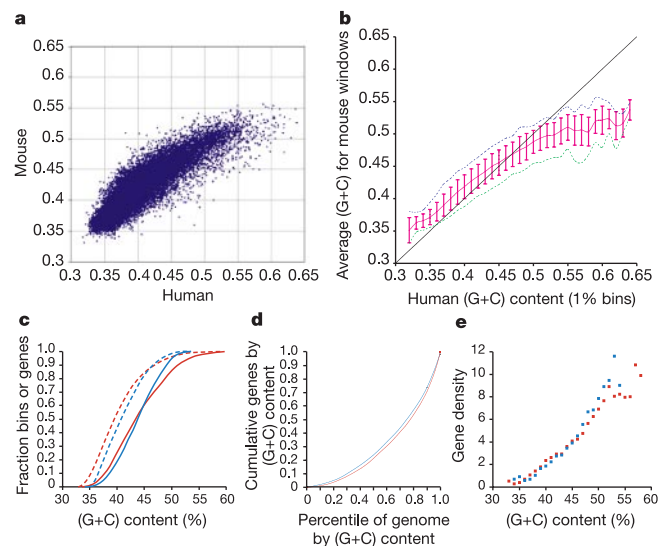


Figure 9 Comparison of (G+C) and gene content in mouse and human. **a**, Scatter plot of mouse (y axis) compared with human (x axis) (G+C) content for all non-overlapping orthologous 100-kb windows. In general, (G+C) content is correlated between the two species, but very few mouse windows have a (G+C) content over 55%, even where the related human window has over 60% (G+C) content. **b**, Average mouse (G+C) content of 100-kb syntenic windows binned by human (G+C) content (1% intervals). The red line indicates median values with standard deviation and 5% (green) and 95% (blue) confidence intervals. The black line indicates identical (G+C) content in orthologous segments. **c–e**, Gene content increases with (G+C) content when comparing (G+C) and gene content in 320-kb non-overlapping, unmasked windows for mouse (blue lines) and human (red lines). **c**, Cumulative proportions of genes (solid lines) and genome (dashed lines) having (G+C) content below a given level. The tighter distribution of (G+C) content in mouse results in the curve for mouse crossing that for human at 45–46% for both genes and total sequence. The tendency for both genomes to be gene-poor at low (G+C) content and gene-rich at high (G+C) content is shown directly in **d**, which shows the fraction of genes residing within the portion of the genome having (G+C) content below a given level (for example, the half of the genome with the lowest (G+C) content contains 25% of the genes). **e**, The average number of genes per window is plotted against the (G+C) content of the window for both genomes, showing that the gene density in mouse reaches the same level as in human but at a lower level of (G+C) content.

Mouse has accumulated more new repetitive sequence than human

Approximately 46% of the human genome can be recognized currently as interspersed repeats resulting from insertions of transposable elements that were active in the last 150–200 million years. The total fraction of the human genome derived from transposons may be considerably larger, but it is not possible to recognize fossils older than a certain age because of the high degree of sequence divergence. Because only 37.5% of the mouse genome is recognized as transposon-derived (Table 5), it is tempting to conclude that the smaller size of the mouse genome is due to lower transposon activity since the divergence of the human and mouse lineages. Closer analysis, however, shows that this is not the case. As we discuss below, transposition has been more active in the mouse lineage. The apparent deficit of transposon-derived sequence in the mouse genome is mostly due to a higher nucleotide substitution rate, which makes it difficult to recognize ancient repeat sequences.

Lineage-specific versus ancestral repeats

Interspersed repeats can be divided into lineage-specific repeats (defined as those introduced by transposition after the divergence of mouse and human) and ancestral repeats (defined as those already present in a common ancestor). Such a division highlights the fact that transposable elements have been more active in the mouse lineage than in the human lineage. Approximately 32.4% of the mouse genome (about 818 Mb) but only 24.4% of the human genome (about 695 Mb) consists of lineage-specific repeats (Table 5). Contrary to initial appearances, transposon insertions have added at least 120 Mb more transposon-derived sequence to the mouse genome than to the human genome since their divergence. This observation is consistent with the previous report that the rate of transposition in the human genome has fallen markedly over the past 40 million years^{1,100}.

The overall lower interspersed repeat density in mouse is the result of an apparent lack of ancestral repeats: they comprise only 5% of the mouse genome compared with 22% of the human genome. The ancestral repeats recognizable in mouse tend to be those of more recent origin, that is, those that originated closest to the mouse–human divergence. This difference may be due partly to a higher deletion rate of non-functional DNA in the mouse lineage, so that more of the older interspersed repeats have been lost. However, the deficit largely reflects a much higher neutral substitution rate in the mouse lineage than in the human lineage, rendering many older ancestral repeats undetectable with available computer programs.

Higher substitution rate in mouse lineage

The hypothesis that the neutral substitution rate is higher in mouse than in human was suggested as early as 1969 (refs 101–103). The idea has continued to be challenged on the basis that the apparent differences may be due to inaccuracies in mammalian phylogenies^{104,105}. The explanation, however, remains unclear, with some attributing it to generation time^{101,106} and others pointing to a closer correlation with body size^{107,108}.

Ancestral repeats provide a powerful measure of neutral substitution rates, on the basis of comparing thousands of current copies to the inferred consensus sequence of the ancestral element. The large copy number and ubiquitous distribution of ancestral repeats overcome issues of local variation in substitution rates (see below). Most notably, differences in divergence levels are not affected by phylogenetic assumptions, as the time spent by an ancestral repeat family in either lineage is necessarily identical.

The median divergence levels of 18 subfamilies of interspersed repeats that were active shortly before the human–rodent speciation (Table 6) indicates an approximately twofold higher average substitution rate in the mouse lineage than in the human lineage, corresponding closely to an early estimate by Wu and Li¹⁰⁹. In human, the least-diverged ancestral repeats have about 16% mis-

Table 5 **Composition of interspersed repeats in the mouse genome**

	Mouse				Human	
	Thousands of copies	Length occupied (Mb)	Fraction of genome (%)	Lineage specific (%)	Fraction of genome (%)	Lineage specific (%)
LINEs	660	475.3	19.20	16.46	20.99	7.94
LINE1	599	464.8	18.78	16.46	17.37	7.94
LINE2	53	9.4	0.38	—	3.30	—
L3/CR1	8	1.2	0.05	—	0.32	—
SINEs	1,498	202.9	8.22	7.63	13.64	10.74
B1 (Alu)	564	67.3	2.66	2.66	10.74	10.74
B2	348	59.6	2.39	2.39	—	—
B4/RSINE	391	57.1	2.36	2.36	—	—
ID	79	5.3	0.25	0.25	—	—
MIR/MIR3	115	14.1	0.57	—	2.90	—
LTR elements	631	244.3	9.87	8.72	8.55	4.09
ERV_classI	34	16.8	0.68	0.58	2.92	2.02
ERV_classII	127	79.1	3.14	3.14	0.30	0.30
ERV_classIII	37	14.0	0.58	0.32	1.55	0.19
MaLRs (III)	388	112.2	4.82	4.02	3.78	1.58
DNA elements	112	21.8	0.88	0.36	3.03	1.00
Charlie	82	15.2	0.62	0.35	1.41	0.14
Other hATs	8	1.6	0.06	—	0.31	—
Tigger	24	4.4	0.17	—	1.06	0.76
Mariner	1	0.2	0.01	0.01	0.10	0.07
Unclassified	26	9.2	0.38	0.37	0.15	0.14
Total	2,926	953.6	38.55	33.53	46.36	24.05
Small RNAs	19	1.5	0.06	0.04	0.04	0.02
Satellites	7	0.7	0.30	NA	0.34	NA
Simple repeats	960	56.1	2.27	NA	0.87	NA

The two right columns show the fractions in the human genome (excluding chromosome Y) for comparison. These and all other interspersed repeat-related data are based on RepeatMasker analysis (version July 2002, sensitive settings, RepBase release 5.3) of the February 2002 mouse and June 2002 human draft assemblies. Each repeat family in the RepeatMasker library was determined to be either order-specific or 'ancestral repeats' present at orthologous sites, usually on the basis of the average divergence level of the interspersed repeat family copies. For elements with an average divergence of 15–19% in human, copies were checked to be present or absent at mouse orthologous sites, to have inserted in known primate-specific repeats, or to have inserts of known mammalian-wide elements. No mammalian-wide repeats were found in the mouse genome that were not already known from the human genome. NA, not applicable.

match to their consensus sequences, which corresponds to approximately 0.17 substitutions per site. In contrast, mouse repeats have diverged by at least 26–27% or about 0.34 substitutions per site, which is about twofold higher than in the human lineage. The total number of substitutions in the two lineages can be estimated at 0.51. Below, we obtain an estimate of a combined rate of 0.46–0.47 substitutions per site, on the basis of an analysis that counts only substitutions since the divergence of the species (see Supplementary Information concerning the methods used).

Assuming a speciation time of 75 Myr, the average substitution rates would have been 2.2×10^{-9} and 4.5×10^{-9} in the human and mouse lineages, respectively. This is in accord with previous

estimates of neutral substitution rates in these organisms. (Reports of highly similar substitution rates in human and mouse lineages relied on a much earlier divergence time of rodents from other mammals¹⁰⁴.)

Comparison of ancestral repeats to their consensus sequence also allows an estimate of the rate of occurrence of small (<50 bp) insertions and deletions (indels). Both species show a net loss of nucleotides (with deleted bases outnumbering inserted bases by at least 2–3-fold), but the overall loss owing to small indels in ancestral repeats is at least twofold higher in mouse than in human. This may contribute a small amount (1–2%) to the difference in genome size noted above.

Table 6 **Divergence levels of interspersed repeats predating the human–mouse speciation**

Interspersed repeat		Mouse				Human				Substitution ratio	Adjusted ratio
		kb	Divergence	Range	JC	kb	Divergence	Range	JC		
L1MA6	LINE1	1,795	0.28	0.04	0.35	2,738	0.16	0.05	0.184	1.92	1.98
L1MA7	LINE1	789	0.28	0.04	0.35	3,502	0.16	0.04	0.181	1.92	1.96
L1MA8	LINE1	951	0.27	0.04	0.34	4,488	0.15	0.04	0.172	1.96	1.96
L1MA9	LINE1	1,032	0.28	0.04	0.35	6,468	0.18	0.05	0.201	1.74	1.86
L1MA10	LINE1	160	0.29	0.04	0.36	1,492	0.19	0.05	0.224	1.61	1.80
L1MB1	LINE1	627	0.29	0.04	0.36	2,947	0.18	0.05	0.211	1.71	1.87
L1MB2	LINE1	725	0.28	0.04	0.35	3,309	0.18	0.06	0.201	1.75	1.87
L1MC1	LINE1	1,389	0.28	0.04	0.36	7,221	0.17	0.05	0.198	1.80	1.92
MLT1A	MaLR	984	0.31	0.04	0.39	2,203	0.21	0.04	0.242	1.62	1.73
MLT1A0	MaLR	1,794	0.30	0.04	0.38	5,424	0.19	0.04	0.219	1.74	1.80
MLT1A1	MaLR	539	0.29	0.04	0.37	1,705	0.19	0.04	0.214	1.74	1.78
MLT1B	MaLR	73	0.28	0.03	0.35	4,482	0.18	0.04	0.203	1.73	1.73
MLT1C	MaLR	2,071	0.30	0.04	0.37	5,511	0.21	0.04	0.245	1.53	1.64
Looper	DNA	33	0.28	0.04	0.34	48	0.18	0.03	0.211	1.62	1.69
MER20	DNA	435	0.29	0.05	0.37	2,205	0.19	0.05	0.222	1.65	1.76
MER33	DNA	232	0.27	0.05	0.33	1,207	0.18	0.04	0.211	1.57	1.63
MER53	DNA	82	0.26	0.05	0.31	524	0.17	0.05	0.191	1.63	1.63
Tigger6a	DNA	97	0.29	0.03	0.37	190	0.18	0.06	0.211	1.77	1.85

Shown are the number of kilobases matched by each subfamily (kb), the median divergence (mismatch) level of all copies from the consensus sequence, the interquartile range of these mismatch levels (range), and a Jukes–Cantor estimate of the substitution level to which the median divergence level corresponds (JC). Notice that RepeatMasker found, on average, four- to fivefold more copies in the human than in the mouse genome, as a result of the higher DNA loss in the rodent lineage as well as a failure to identify many highly diverged copies. The two right columns contain the ratio of the JC substitution level in mouse over human, and an adjusted ratio (AR) of the mouse and human substitution level after subtraction from both of the approximate fraction accumulated in the common human–mouse ancestor. For this fraction we have taken the difference between the ancestral repeat average substitution level and least diverged ancestral repeat family (L1MA8). See the Supplementary Information for a discussion of the origin of the variance in the human and mouse ratios.

It should be noted that the roughly twofold higher substitution rate in mouse represents an average rate since the time of divergence, including an initial period when the two lineages had comparable rates. Comparison with more recent relatives (mouse–rat and human–gibbon, each about 20–25 Myr) indicate that the current substitution rate per year in mouse is probably much higher, perhaps about fivefold higher (see Supplementary Information). Also, note that these estimates refer to substitution rate per year, rather than per generation. Because the human generation time is much longer than that of the mouse (by at least 20-fold), the substitution rate is greater in human than mouse when measured per generation.

Higher substitution rate obscures old repeats

We measured the impact of the higher substitution rate in mouse on the ability to detect ancestral repeats in the mouse genome. By computer simulation, the ability of the RepeatMasker¹⁰⁰ program to detect repeats was found to fall off rapidly for divergence levels above about 37%. If we simulate the events in the mouse lineage by adjusting the ancestral repeats in the human genome for the higher substitution levels that would have occurred in the mouse genome, the proportion of the genome that would still be recognizable as ancestral repeats falls to only 6%. This is in close agreement with the proportion actually observed for the mouse. Thus, the current analysis of repeated sequences allows us to see further back into human history (roughly 150–200 Myr) than into mouse history (roughly 100–120 Myr).

A higher rate of interspersed repeat insertion does not explain the larger size of the human genome. Below, we suggest that the explanation lies in a higher rate of large deletions in the mouse lineage.

Comparison of mouse and human repeats

All mammals have essentially the same four classes of transposable elements: (1) the autonomous long interspersed nucleotide element (LINE)-like elements; (2) the LINE-dependent, short RNA-derived short interspersed nucleotide elements (SINEs); (3) retrovirus-like elements with long terminal repeats (LTRs); and (4) DNA transposons. The first three classes procreate by reverse transcription of an RNA intermediate (retroposition), whereas DNA transposons move by a cut-and-paste mechanism of DNA sequence (see refs 1, 100 for further information about these classes).

A comparison of these repeat classes in the mouse and human genomes can be enlightening. On the one hand, differences between the two species reveal the dynamic nature of transposable elements; on the other hand, similarities in the location of lineage-specific elements point to common biological factors that govern insertion and retention of interspersed repeats.

Differences between mouse and human

The most notable difference is in the changing rate of transposition over time: the rate has remained fairly constant in mouse, but markedly increased to a peak at about 40 Myr in human, and then plummeted. This phenomenon was noted in our initial analysis of

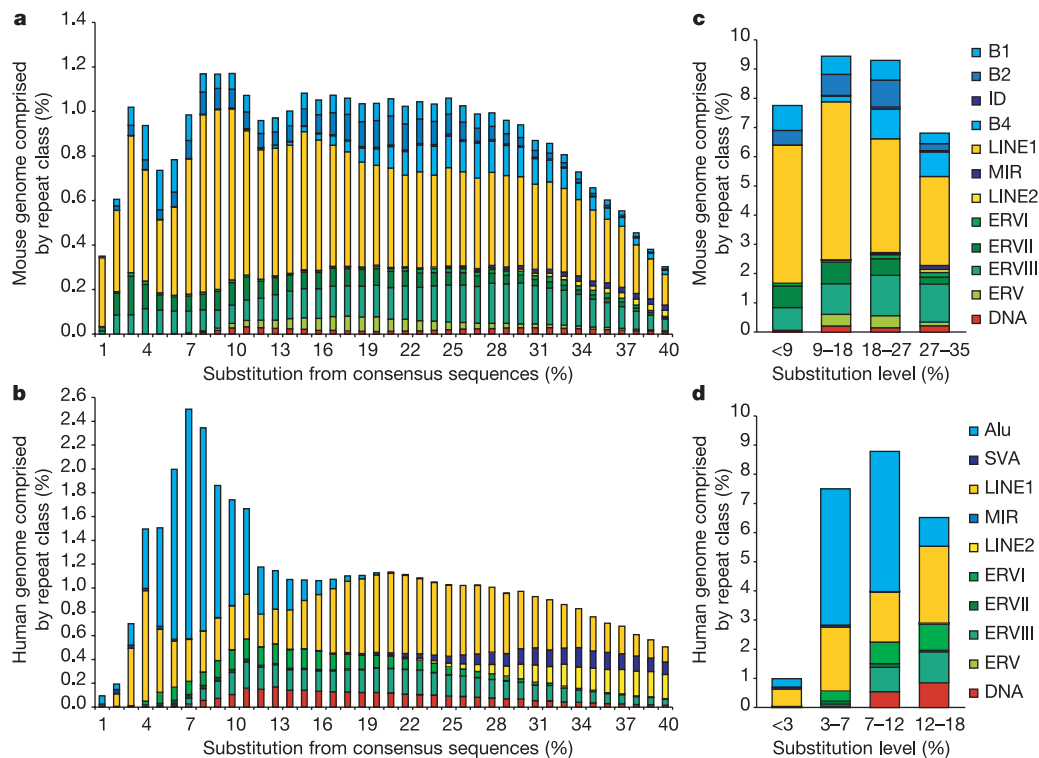


Figure 10 Age distribution of interspersed repeats in the mouse and human genomes. This is an update of Fig. 18 in the IHGC human genome paper¹. **a, b**, Distribution for mouse and human of copies of each repeat class in bins corresponding to 1% increments in substitution level calculated using Jukes–Cantor formula ($K = -3/4 \ln(1 - D_{\text{rest}}^*/4/3)$) (see Supplementary Information for definition). The first bin for mouse is artificially low because the WGS assembly used for mouse excludes a larger percentage of very recent repeats. **c, d**, Interspersed repeats grouped into bins of approximately equal time periods after adjusting for the different rates of substitution in the two genomes. On average, the substitution level has been twofold higher in the mouse than in the human lineage (Table

6), but the difference was initially less and has increased over time. The present rates may differ over fourfold. The activity of transposable elements in the mouse lineage has been quite uniform compared with the human lineage, where an overall decline was interrupted temporarily by a burst of Alu activity. The apparent absence of <2% diverged interspersed repeats in mouse is primarily due to the shotgun sequencing strategy; long, closely similar interspersed repeats very often were not assembled. This is supported by an up to tenfold higher concentration of young L1 and ERV elements at the edges of gaps. The gradually decreasing density of repeats beyond a 30% substitution level reflects in part the limits of the detection method.

the human genome; the availability of the mouse genome sequence now confirms and sharpens the observation (Fig. 10). Beyond this overall tendency, there are specific differences in each of the four repeat classes.

The first class that we discuss is LINEs. Copies of LINE1 (L1) form the single largest fraction of interspersed repeat sequence in both human and mouse. No other LINE seems to have been active in either lineage. The extant L1 elements in both species derive from a common ancestor (L1MA6 in Table 6) by means of a series of subfamilies defined primarily by the rapidly evolving 3' non-coding sequences¹¹⁰. The L1 5'-untranslated regions (UTRs) in both lineages have been even more variable, occasionally through acquisition of entirely new sequences¹¹¹. Indeed, the three active subfamilies in mouse, which are otherwise >97% identical, have unrelated or highly diverged 5' ends^{112–114}. L1 seems to have remained highly active in mouse, whereas it has declined in the human lineage. Goodier and co-workers¹¹³ estimated that the mouse genome contains at least 3,000 potentially active elements (full-length with two intact open reading frames (ORFs)). The current draft sequence of the mouse genome contains only 400 young, full-length elements; of these only 12 have two intact ORFs. This is probably a reflection of the WGS shotgun approach used to assemble the genome. Indeed, most of the young elements in the draft genome sequence are incomplete owing to internal sequence gaps, reflecting the difficulty that WGS assembly has with highly similar repeat sequences. This is a notable limitation of the draft sequence.

The second repeat class is SINEs. Whereas only a single SINE (Alu) was active in the human lineage, the mouse lineage has been exposed to four distinct SINEs (B1, B2, ID, B4). Each is thought to rely on L1 for retroposition, although none share sequence similarity, as is the rule for other LINE–SINE pairs^{115,116}. The mouse B1 and human Alu SINEs are unique among known SINEs in being derived from 7SL RNA; they probably have a common origin¹¹⁷. The mouse B2 is typical among SINEs in having a transfer RNA-derived promoter region. Recent ID elements seem to be derived from a neuronally expressed RNA gene called *BCI*, which may itself have been recruited from an earlier SINE. This subfamily is minor in mouse, with 2–4,000 copies, but has expanded rapidly in rat where it has produced more than 130,000 copies since the mouse–rat speciation¹¹⁸. Both B2 and ID closely resemble Ala-tRNA, but seem to have independent origins. The B4 family resembles a fusion between B1 and ID^{119,120}. We found that 25% of the 75,000 identified ID elements were located within 50 bp of a B1 element of similar orientation, suggesting that perhaps most older ID elements are mislabelled or truncated B4 SINEs.

More rodent-specific SINEs are present in the mouse genome than Alu SINEs in human (1.4 and 1.1 million, respectively), but they occupy a smaller portion of the genome (7.6% and 10.7%, respectively) because of their smaller sizes. The existence of four families in mouse provides independent opportunities to investigate the properties of SINEs (see below).

The third repeat class is LTR elements. All interspersed LTR-containing elements in mammals are derivatives of the vertebrate-specific retrovirus clade of retrotransposons. The earliest infectious retroviruses probably originated from endogenous retroviral-like (ERV) elements that acquired mechanisms for horizontal transmission¹²¹, whereas many current endogenous retroviral elements have probably arisen from infection by retroviruses.

Endogenous retroviruses fall into three classes (I–III), which show a markedly dissimilar evolutionary history in human and mouse (see Fig. 10). Notably, ERVs are nearly extinct in human whereas all three classes have active members in mouse.

Class III accounts for 80% of recognized LTR element copies predating the human–mouse speciation. This class includes the non-autonomous MaLRs: with 388,000 recognizable copies in mouse, it is the single most successful LTR element. It is still active

in mouse (represented by MERVL and the MT and ORR1 MaLRs), but died out some 50 Myr in human¹²².

Copies of class II elements are tenfold denser in mouse than in human. Among the active class II elements in mouse are two abundant and active groups, the intracisternal-A particles (IAP) and the early-transposons (ETn). About 15% of all spontaneous mouse mutants have an allele associated with IAP or ETn insertion, demonstrating the functional consequences of class I element activity in mice. A third active class, the mouse mammary tumour virus, is present in only a few copies¹²³ (see Supplementary Information). In human, there is evidence for at most a few active elements (HERVK10 and HERBK113 (ref. 124)). No class II ERVs are known to predate the human–mouse speciation.

In contrast, class I element copies are fourfold more common in the human than the mouse genome (although it is possible that some have not yet been recognized in mouse). In mouse, this class includes active ERVs, such as the murine leukaemia virus, MuRRS, MuRVY and VL30 (several of which have caused insertional mutations in mouse)—no similar activity is known to exist in human. It is unclear why the class I ERVs have been more successful in the human lineage whereas the class II ERVs have flourished in the mouse lineage.

The fourth repeat class is the DNA transposons. Although most transposable elements have been more active in mouse than human, DNA transposons show the reverse pattern. Only four lineage-specific DNA transposon families could be identified in mouse (the mariner element MMAR1, and the hAT elements URR1, RMER30 and RChar1), compared with 14 in the primate lineage.

For evolutionary survival, DNA transposons are thought to depend on frequent horizontal transfer to new host genomes by means of vectors such as viruses and other intracellular parasites^{116,125}. The mammalian immune system probably forms a large obstacle to the successful invasion of DNA transposons. Perhaps the rodent germ line has been harder to infiltrate by horizontal transfer than the primate genome. Alternatively, it is possible that highly diverged families active in early rodent evolution have not been detected yet. Notably, most copies in the human genome were deposited early in primate evolution.

An interesting case is the mariner element, which seems to have infiltrated independently both the rodent and human genomes. The mariner element is represented by elements (MMAR1 in mouse and HSMAR1 in human) that are 97% identical. The average substitution level outside CpG sites of HSMAR1 is 8% and of MMAR1 is 22%, both well below the divergence of elements predating the human–mouse speciation (Table 6).

Some of the above differences in the nature of interspersed repeats in human and mouse could reflect systematic factors in mouse and human biology, whereas others may represent random fluctuations. Deeper understanding of the biology of transposable elements and detailed knowledge of interspersed repeat populations in other mammals should clarify these issues.

Similar repeats accumulate in orthologous locations

One of the most notable features about repeat elements is the contrast in the genomic distribution of LINEs and SINEs. Whereas LINEs are strongly biased towards (A+T)-rich regions, SINEs are strongly biased towards (G+C)-rich regions. The contrast is all the more notable because both elements are inserted into the genome through the action of the same endonuclease^{126,127}.

Such preferences were studied in detail in the initial analysis of the human genome¹, and essentially equivalent preferences are seen in the mouse genome (Fig. 11). With the availability of two mammalian genomes, however, it is possible to extend this analysis to explore whether (A+T) and (G+C) content are truly causative factors or merely reflections of an underlying biological process.

Towards that end, we studied the insertion of lineage-specific repeat elements in orthologous segments in the human and mouse

genomes (Fig. 12). Each insertion represents a new, independent event occurring in one lineage, and thus any correlation between the two species reflects underlying proclivity to insert or retain repeats in particular regions. Visual inspection reveals a strong correlation in the sites of lineage-specific repeats of the various classes (Fig. 12). Lineage-specific repeats also correlate with other genomic features, as discussed in the section on genome evolution.

The correlation of local lineage-specific SINE density is extremely strong (Fig. 13a). Moreover, local SINE density in one species is better predicted by SINE density in the other species than it is by local (G+C) content (Table 7). The local density of each distinct rodent-specific type of SINE is a strong predictor of Alu density at the orthologous locus in human, although the Alu equivalent B1 SINEs show the strongest correlation ($r^2 = 0.784$) (Table 7).

We interpret these results to mean that SINE density is influenced by genomic features that are correlated with (G+C) content but that are distinct from (G+C) content *per se*. The fact that (G+C) content alone does not determine SINE density is consistent with the observation that some (G+C)-rich regions of the human genome are not Alu rich^{128,129}.

Lineage-specific LINE density is also clearly correlated between mouse and human (Fig. 13b), although the relationship does not seem to be linear and it is not as strong (Spearman rank analysis, $r^2 = 0.45$). (G+C) content seems to contribute as an independent variable (increasing r^2 to 0.52), suggesting that (G+C) content itself directly affects LINE integration.

Genomic outliers

In addition to examining the general correlation in repeat density between mouse and human, we also considered some of the extreme examples. In the human genome, the four homeobox clusters (*HOXA*, *HOXB*, *HOXC* and *HOXD*) are by far the most repeat-poor regions of the human genome, with repeat content in the range

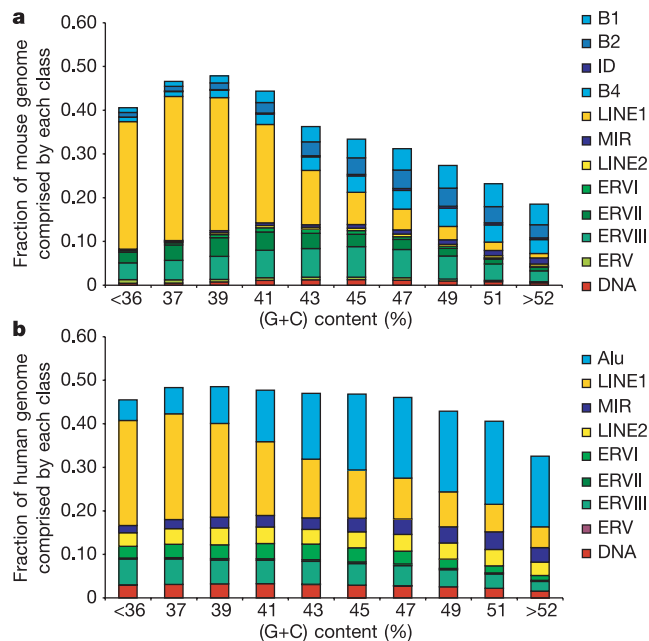


Figure 11 Density of interspersed repeat classes at different (G+C) content in the mouse (a) and human (b) genomes. In both species, there is a strong increase in SINE density and a decrease in L1 density with increasing (G+C) content, with the latter particularly marked in the mouse. Another notable contrast is that in mouse, overall interspersed repeat density gradually decreases 2.5-fold with increasing (G+C) content, whereas in human the overall repeat density remains quite uniform. This reflects both the abundance of L1 elements in the mouse (G+C)-poor regions and the unusually high density of Alu in human (G+C)-rich regions.

of 1%. These same four regions are exceptions in the mouse genome as well. The strong selective constraints against insertion in these regions probably reflect dense, long-range regulatory information across this developmentally important gene cluster. Other repeat-poor loci in the human genome¹ (about 100-kb regions on human chromosomes 1p36, 8q21 and 18q22) have independently remained repeat-poor in mouse (3.6, 6.5 and 7%, respectively) over roughly 75 million years of evolution; we speculate that this similarly reflects dense regulatory information in the region.

Conversely, we searched the mouse genome for repeat-poor regions of at least 100 kb. Again, the outliers show a clear tendency to be repeat-poor in human (see Supplementary Information). A notable feature is that in half of the selected loci the repeat-poor region is confined almost exactly to the extent of a single gene. Figure 14 shows this for the *Zfx1b* locus, and also shows coincidence of exclusion of interspersed repeats with high conservation between human and mouse.

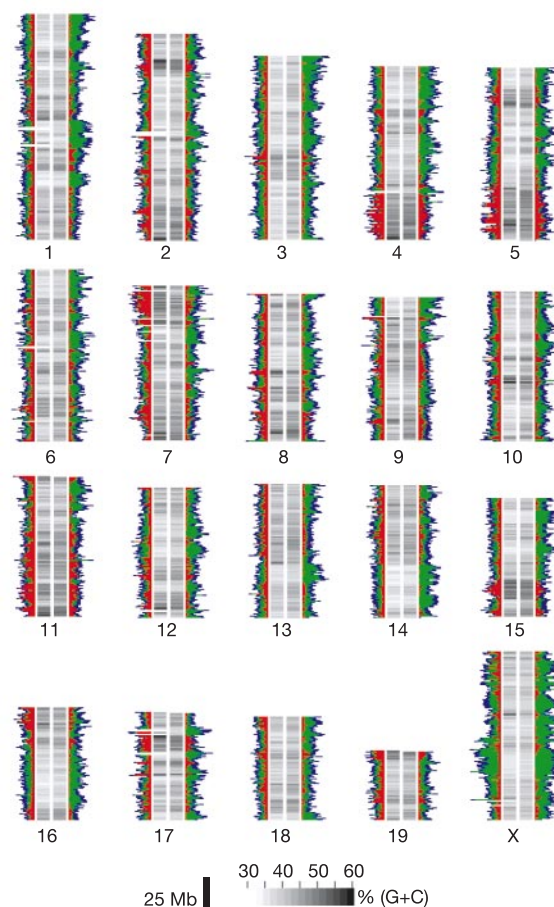


Figure 12 Conservation of (G+C) content and convergence of interspersed repeat distribution between the human and mouse genomes. For each mouse chromosome, its (G+C) content is depicted as a greyscale (centre, right), with darker shades indicating (G+C)-richer regions. Rodent-specific repeats are shown as cumulative histograms (far right), with red, green and blue indicating SINEs, LINEs and other repeats, respectively. The (G+C) content of the orthologous human sequence is similarly shown (centre, left) as well as the primate-specific repeats (far left). Gaps in the human sequence appear opposite those regions of the mouse genome lacking assigned conserved syntenic segments. Note the correlation in (G+C) and repeat content between orthologous regions of the two genomes. Many abrupt shifts in (G+C) content and repeat density are clearly associated with syntenic breaks, which are therefore more likely to be breaks associated with the rodent lineage⁴⁵.

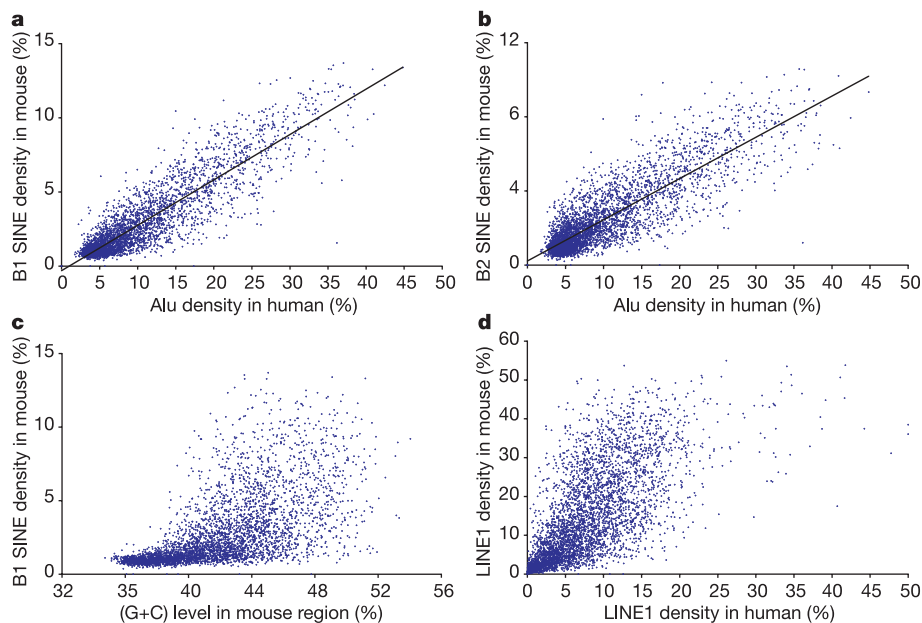


Figure 13 Correlation of order-specific SINEs and LINEs in human and mouse orthologous regions. SINE and LINE densities were calculated for 4,126 orthologous pairs with a constant size of 500 kb in mouse. **a, b**, Strong linear correlation of Alu density in human, and both the Alu-like B1 SINEs (**a**) and the unrelated B2 SINEs (**b**) densities in mouse.

These correlations are stronger than the correlation of SINE density with (G+C) level (**c**). **d**, The relationship of LINE1 density in human and mouse orthologous regions is not linear, reflecting the more extreme bias of LINE1 for (A+T)-rich DNA in mouse.

LINE elements prefer sex chromosomes

A conspicuous feature of the repeat distribution is that LINE elements in both human and mouse show a preference for accumulating on sex chromosomes (Figs 12 and 15). Mouse chromosome X contains almost twice the density of lineage-specific L1 copies as the mouse autosomes (28.5% compared with 14.6%). Human sex chromosomes show an even stronger bias (17.5% on X and 18.0% on Y compared with 7.5% for the autosomes). The enrichment is still highly significant even after accounting for the generally higher (A+T) content of the sex chromosomes (Fig. 15). The higher density of L1 on sex chromosomes had been noted in early hybridization experiments^{130,131} and has led to the suggestion that L1 copies may help facilitate X inactivation^{132,133}. For chromosome Y, the accumulation probably reflects a greater tolerance for insertion (owing to the paucity of genes) and the inability to purge deleterious mutations by recombination. Consistent with the latter explanation, chromosome Y also shows a threefold higher density of full-length L1 copies (which are rapidly eliminated elsewhere in the genome¹³⁴) and an overall excess of LTR element insertions.

Chromosome X shows an excess of L1 copies, but not a marked excess of either full-length L1 or LTR copies. The explanation for this preferential accumulation of L1 elements on chromosome X in both the mouse and human lineages remains unclear.

Simple sequence repeats

Mammalian genomes are scattered with simple sequence repeats (SSRs), consisting of short perfect or near-perfect tandem repeats that presumably arise through slippage during DNA replication. SSRs have had a particularly important role as genetic markers in linkage studies in both mouse and human, because their lengths tend to be polymorphic in populations and can be readily assayed by PCR. It is possible that such SSRs, arising as they do through replication errors, would be largely equivalent between mouse and human; however, there are impressive differences between the two species¹³⁵.

Table 7 Predicted repeat density in human based on mouse		
Predicted density of repeat in human	Linear regression	Spearman rank
Alu density based on:		
Density in orthologous sites for all SINEs	0.79	0.68
Density in orthologous sites for B1	0.81	0.70
Density in orthologous sites for B2	0.73	0.62
Density in orthologous sites for ID + B4	0.49	0.54
SINE density in mouse DNA of equivalent (G+C) content	0.21	0.31
Local (G+C) content	0.40	0.48
Density in orthologous sites for all SINEs and local (G+C) content	0.80	0.70
LINE1 density based on:		
LINE1 density at orthologous mouse regions	0.45	0.54
LINE1 density in mouse DNA of equivalent (G+C) content	0.26	0.42
Local (G+C) content	0.37	0.51
LINE1 density at orthologous mouse regions and local (G+C) content	0.49	0.59

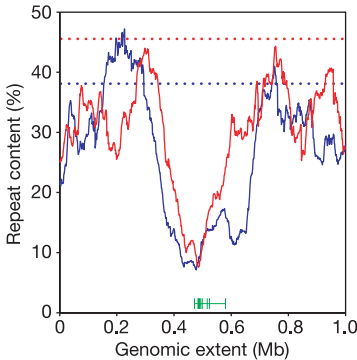


Figure 14 The zinc-finger homeobox 1b (*Zfx1b*) loci in human and mouse are both repeat poor. The repeat content for mouse (blue) and human (red) in 50-kb windows is shown for a 1-Mb region surrounding the *Zfx1b* gene (green). Dotted lines indicate genome average for repeat content in mouse (blue) and human (red). The repeat-poor regions (<10% repeat content in mouse and human) coincide with the location of the 150-kb-long gene and regions of high conservation between human and mouse.

Overall, mouse has 2.25–3.25-fold more short SSRs (1–5 bp unit) than human (Table 8); the precise ratio depends on the percentage identity required in defining a tandem repeat. The mouse seems to represent an exception among mammals on the basis of comparison with the small amount of genomic sequence available from dog (4 Mb) and pig (5 Mb), both of which show proportions closer to human¹³⁶ (E. Green, unpublished data; Table 8).

The analysis can be refined, however, by excluding transposable elements that contain SSRs at their 3' ends. For example, 90% of A-rich SSRs in human are provided by or spawned from poly(A) tails of Alu and L1 elements, and 15% of (CA)_n-like SSRs in mouse are contained in B2 element tails. When these sources are eliminated, the contrast between mouse and human grows to roughly fourfold.

The reason for the greater density of SSRs in mouse is unknown. Table 9 shows that SSRs of >20 bp are not only more frequent, but are generally also longer in the mouse than in the human genome, suggesting that this difference is due to extension rather than to initiation. The equilibrium distribution of SSR length has been proposed¹³⁷ to be determined by slippage between exact copies of the repeat during meiotic recombination¹³⁸. The shorter lengths of SSRs in human may result from the higher rate of point substitutions per generation (see above), which disrupts the exactness of the repeats.

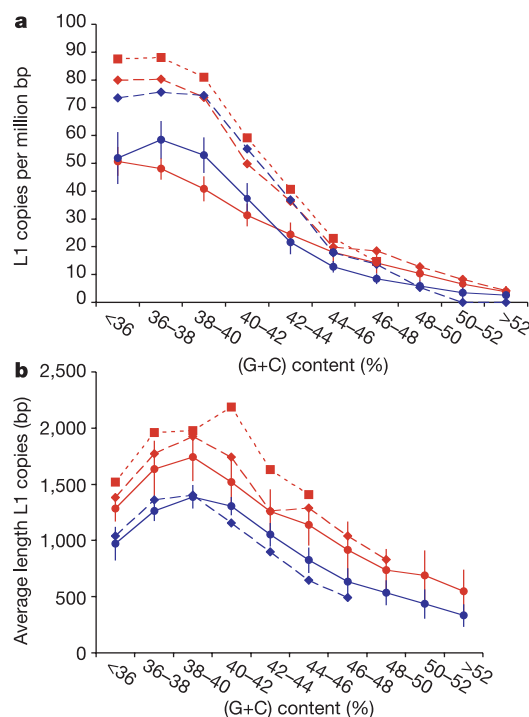


Figure 15 Comparison of L1 characteristics of autosomes and sex chromosomes as a function of (G+C) content in mouse (blue) and human (red). Error bars depict standard deviation over all autosomes (circles). Diamonds, X chromosomes; squares, human Y chromosome. The mouse Y chromosome is not represented in the whole-genome assembly, and too little clone-based information is available to be included. **a**, The number of lineage-specific L1 copies per megabase declines 13- to 20-fold from lowest to highest (G+C) content. This relationship is stronger in mouse and on the sex chromosomes. Note that, for the same (G+C) content, L1 density is 1.5- to twofold higher on the sex chromosomes. **b**, The average length of lineage-specific L1 copies peaks at around the 39% (G+C) level, where it is three- (human) to fourfold (mouse) higher than in the (G+C)-richest regions. The average length in mouse is underestimated owing to the bias against full-length young elements in the shotgun assembly. On average, L1 copies are longer on human Y than on either X chromosome or the autosomes.

Apart from the absolute number of SSRs, there are also some marked differences in the frequency of certain SSR classes (Table 9)¹³⁶. The most extreme is the tetramer (ACAG)_n, which is 20-fold more common in mouse than human (even after eliminating copies associated with B2 and B4 SINEs); the sequence does not occur in large clusters, but rather is distributed throughout the genome. In general, SSRs in which one strand is a polypurine tract and the other a polypyrimidine tract are much more common and extended in mouse than human. For the six such di-, tri- and tetramer SSRs (AG, AAG, AGG, AAAG, AAGG, AGGG), copies with at least 20 bp and 95% identity are 1.6-fold longer and tenfold more common in mouse than human.

Analysis of the distribution of SSRs across chromosomes also reveals an interesting feature common to both organisms (see Supplementary Information). In both human and mouse, there is a nearly twofold increase in density of SSRs near the distal ends of chromosome arms. Because mouse chromosomes are acrocentric, they show the effect only at one end. The increased density of SSRs in telomeric regions may reflect the tendency towards higher recombination rates in subtelomeric regions¹.

Mouse genes

Genes comprise only a small portion of the mammalian genome, but they are understandably the focus of greatest interest. One of the most notable findings of the initial sequencing and analysis of the human genome¹ was that the number of protein-coding genes was only in the range of 30,000–40,000, far less than the widely cited textbook figure of 100,000, but in accord with more recent, rigorous estimates^{55,139–141}. The lower gene count was based on the observed and predicted gene counts, statistically adjusted for systematic under- and overcounting.

Our goal here is to produce an improved catalogue of mammalian protein-coding genes and to revisit the gene count. Genome analysis has been enhanced by a number of recent developments. These include burgeoning mammalian EST and cDNA collections, knowledge of the genomes and proteomes of a growing number of organisms, increasingly complete coverage of the mouse and human genomes in high-quality sequence assemblies, and the ability to use *de novo* gene prediction methodologies that exploit information from two mammalian genomes to avoid potential biases inherent in using known transcripts or homology to known genes.

We focus here on protein-coding genes, because the ability to recognize new RNA genes remains rudimentary. As used below, the terms 'gene catalogue' and 'gene count' refer to protein-coding genes only. We briefly discuss RNA genes at the end of the section.

Evidence-based gene prediction

We constructed catalogues of human and mouse gene predictions on the basis of available experimental evidence. The main computational tool was the Ensembl gene prediction pipeline¹⁴² augmented with the Genie gene prediction pipeline¹⁴³. Briefly, the Ensembl

Table 8 Density of short SSRs in mouse compared with other mammals

Cutoff (%)	All 1–5-bp unit SSRs including those in IRs					Excluding SSRs within IRs and IR tails		
	Mouse (%)	Human (%)	Ratio*	Dog (%)	Pig (%)	Mouse (%)	Human (%)	Ratio*
5	0.82	0.25	3.24	0.38	0.25	0.64	0.15	4.41
10	1.35	0.46	2.91	0.72	0.50	1.03	0.25	4.19
15	1.86	0.68	2.73	1.14	0.72	1.38	0.37	3.77
None	2.67	1.19	2.24	1.90	1.27	1.99	0.64	3.10

The cutoff indicates the maximum level of imperfect copies allowed, with 'none' indicating that all SSRs are recognized by RepeatMasker. To determine which SSRs were spawned from within IRs or IR tails, the locations of SSRs were overlapped with the RepeatMasker output. We excluded those SSRs overlapped by IRs and those resembling unmasked tails; that is, occurring more than one-third of the time adjacent to the same type of IR. IR, interspersed repeat; SSR, simple sequence repeat.

*Ratio was determined by dividing mouse percentage by human percentage.

Table 9 Frequency of different SSRs in mouse and human

Simple sequence repeat (SSR)	Including SSRs in IRs				Excluding SSRs from SINE and LINE tails and within IRs		
	Fraction of mouse genome (bp Mb ⁻¹)	Average no. units per SSR (mouse)	Average no. units per SSR (human)	Frequency (number SSRs per Mb)	Frequency ratio (mouse/human)	Frequency (number per Mb)	Frequency ratio (mouse/human)
A	555	26.7	25.0	20.8	0.4	10.4	1.6
C	11	22.7	25.2	0.5	6.2	0.2	4.6
AC	3070	20.8	18.1	73.8	3.5	62.1	3.2
AG	1365	24.3	15.7	28.1	6.9	26.7	7.5
AT	370	21.2	17.8	8.7	1.9	8.6	2.2
CG	4	11.6	11.2	0.2	10.5	0.1	11.1
AAC	148	9.5	8.6	5.2	1.7	3.9	2.8
AAG	152	29.1	15.9	1.7	11.7	1.6	14.5
AAT	147	12.2	9.7	4.0	0.9	2.9	1.9
ACC	53	11.2	9.6	1.6	6.2	1.3	6.4
AGC	30	11.3	8.8	0.9	3.0	0.8	3.3
AGG	46	12.2	8.9	1.3	5.1	1.1	6.4
AAAC	465	6.7	6.2	17.5	2.2	10.3	4.5
AAAG	203	16.2	10.2	3.1	2.8	2.3	5.6
AAAT	346	8.2	7.1	10.5	0.8	5.8	2.0
AACC	36	8.8	7.2	1.0	8.3	0.6	6.1
AAGG	122	13.5	12.3	2.3	4.9	2.0	6.7
AATG	41	7.8	6.7	1.3	1.1	1.0	1.6
ACAG	70	7.4	6.5	2.4	21.1	1.8	21.2
ACAT	85	10.3	8.4	2.1	6.0	1.7	8.6
AGAT	282	16.2	12.8	4.4	4.0	3.7	4.0
AGGG	39	8.2	6.6	1.2	6.2	1.1	8.8
AAAAAC	369	5.9	5.0	12.4	1.5	7.2	2.9

Frequency and density of monomeric and dimeric SSRs and the most common trimeric and tetrameric SSRs. The numbers are for >20-bp-long SSRs containing <10% substitutions and indels. The two right columns show the density of each repeat, excluding those SSRs spawned from the tails of SINEs and LINEs or inherently part of other IRs. For this, SSRs were counted in a default (-m -s) RepeatMasker run.

system uses three tiers of input. First, known protein-coding cDNAs are mapped onto the genome. Second, additional protein-coding genes are predicted on the basis of similarity to proteins in any organism using the GeneWise program¹⁴⁴. Third, *de novo* gene predictions from the GENSCAN program¹⁴⁵ that are supported by experimental evidence (such as ESTs) are considered. These three strands of evidence are reconciled into a single gene catalogue by using heuristics to merge overlapping predictions, detect pseudogenes and discard misassemblies. These results are then augmented by using conservative predictions from the Genie system, which predicts gene structures in the genomic regions delimited by paired 5' and 3' ESTs on the basis of cDNA and EST information from the region.

We also examined predictions from a variety of other computational systems (see Supplementary Information). These methods tended to have significant overlap with the above-generated gene catalogues, but each tended to introduce significant numbers of predictions that were unsupported by other methods and that appeared to be false positives. Accordingly, we did not add these predictions to our gene catalogues; however, we did use them to fill in missing exons in existing predictions (see Supplementary Information).

The computational pipeline produces predicted transcripts, which may represent fragmentary products or alternative products of a gene. They may also represent pseudogenes, which can be difficult in some cases to distinguish from real genes. The predicted transcripts are then aggregated into predicted genes on the basis of sequence overlaps (see Supplementary Information). The computational pipeline remains imperfect and the predictions are tentative.

Initial and current human gene catalogue

The initial human gene catalogue¹ contained about 45,000 predicted transcripts, which were aggregated into about 32,000 predicted genes containing a total of approximately 170,000 distinct exons (Table 10). Many of the predicted transcripts clearly represented only gene fragments, because the overall set contained considerably fewer exons per gene (mean 4.3, median 3) than

known full-length human genes (mean 10.2, median 8).

This initial gene catalogue was used to estimate the number of human protein-coding genes, on the basis of estimates of the fragmentation rate, false positive rate and false negative rate for true human genes. Such corrections were particularly important, because a typical human gene was represented in the predictions by about half of its coding sequence or was significantly fragmented. The analysis suggested that the roughly 32,000 predicted genes represented about 24,500 actual human genes (on the basis of fragmentation and false positive rates) out of the best-estimate total of approximately 31,000 human protein-coding genes on the basis of estimated false negatives¹. We suggested a range of 30,000–40,000 to allow for additional genes.

Several papers have re-analysed the initial gene catalogue and argued for a substantially larger human gene count^{146,147}. Most of these analyses, however, did not account for the incomplete nature of the catalogue¹⁴⁸, the complexities arising from alternative splicing, and the difficulty of interpreting evidence from fragmentary messenger RNAs (such as ESTs and serial analysis of gene expression (SAGE) tags) that may not represent protein-coding genes¹⁴⁹.

Since the initial paper¹, the human gene catalogue has been refined as sequence becomes more complete and methods are

Table 10 Gene count in human and mouse genomes

Genome feature	Human		Mouse	
	Initial (Feb. 2001)	Current (Sept. 2002)	Initial* (this paper)	Extended† (this paper)
Predicted transcripts	44,860	27,048	28,097	29,201
Predicted genes	31,778	22,808	22,444	22,011
Known cDNAs	14,882	17,152	13,591	12,226
New predictions	16,896	5,656	8,853	9,785
Mean exons/transcript‡	4.2 (3)	8.7 (6)	8.2 (6)	8.4 (6)
Total predicted exons	170,211	198,889	191,290	213,562

*Without RIKEN cDNA set.

†With RIKEN cDNA set.

‡Median values are in parentheses.

revised. The current catalogue (Ensembl build 29) contains 27,049 predicted transcripts aggregated into 22,808 predicted genes containing about 199,000 distinct exons (Table 10). The predicted transcripts are larger, with the mean number of exons roughly doubling (to 8.7), and the catalogue has increased in completeness, with the total number of exons increasing by nearly 20%. We return below to the issue of estimating the mammalian gene count.

Mouse gene catalogue

We sought to create a mouse gene catalogue using the same methodology as that used for the human gene catalogue (Table 10). An initial catalogue was created by using the same evidence set as for the human analysis, including cDNAs and proteins from various organisms. This set included a previously published collection of mouse cDNAs produced at the RIKEN Genome Center⁴¹.

We also created an extended mouse gene catalogue by including a much larger set of about 32,000 mouse cDNAs with significant ORFs (see Supplementary Information) that were sequenced by RIKEN (see ref. 150). These additional mouse cDNAs improved the catalogue by increasing the average transcript length through the addition of exons (raising the total from about 191,000 to about 213,000, including many from untranslated regions) and by joining fragmented transcripts. The set contributed roughly 1,200 new predicted genes. The total number of predicted genes did not change significantly, however, because the increase was offset by a decrease due to mergers of predicted genes. These mouse cDNAs have not yet been used to extend the human gene catalogue. Accordingly, comparisons of the mouse and human gene catalogues below use the initial mouse gene catalogue.

The extended mouse gene catalogue contains 29,201 predicted transcripts, corresponding to 22,011 predicted genes that contain about 213,500 distinct exons. These include 12,226 transcripts corresponding to cDNAs in the public databases, with 7,481 of these in the well-curated RefSeq collection¹⁵¹. There are 9,785 predicted transcripts that do not correspond to known cDNAs, but these are built on the basis of similarity to known proteins.

The new mouse and human gene catalogues contain many new genes not previously identified in either genome. These include new paralogues for genes responsible for at least five diseases: *RFX5*, responsible for a type of severe combined immunodeficiency resulting from lack of expression of human leukocyte antigen (HLA) antigens on certain haematopoietic cells¹⁵²; *bestrophin*, responsible for a form of muscular degeneration¹⁵³; *otoferlin*, responsible for a non-syndromic prelingual deafness¹⁵⁴; *Crumbs1*, mutated in two inherited eye disorders^{155,156}; and *adiponectin*, a deficiency of which leads to diet-induced insulin resistance in mice¹⁵⁷. The *RFX5* case is interesting, because disruption of the known mouse homologue alone does not reproduce the human disease, but may do so in conjunction with disruption of the newly identified paralogue¹⁵⁸.

Recently, Mural and colleagues⁴⁵ analysed the sequence of mouse chromosome 16 and reported 731 gene predictions (compared with 756 gene predictions in our set for chromosome 16). Our gene catalogue contains 656 of these gene predictions, indicating extensive agreement between these two independent analyses. Most of the remaining 75 genes reported by ref. 45 seem to be systematic errors (common to all such programs), such as relatively short gene predictions arising from protein matches to low-complexity regions.

It should be emphasized that the human and mouse gene catalogues, although increasingly complete, remain imperfect. Both genome sequences are still incomplete. Some authentic genes are missing, fragmented or otherwise incorrectly described, and some predicted genes are pseudogenes or are otherwise spurious. We describe below further analysis of these challenges.

Pseudogenes

An important issue in annotating mammalian genomes is distinguishing real genes from pseudogenes, that is, inactive gene copies. Processed pseudogenes arise through retrotransposition of spliced or partially spliced mRNA into the genome; they are often recognized by the loss of some or all introns relative to other copies of the gene. Unprocessed pseudogenes arise from duplication of genomic regions or from the degeneration of an extant gene that has been released from selection. They sometimes contain all exons, but often have suffered deletions and rearrangements that may make it difficult to recognize their precise parentage. Over time, pseudogenes of either class tend to accumulate mutations that clearly reveal them to be inactive, such as multiple frameshifts or stop codons. More generally, they acquire a larger ratio of non-synonymous to synonymous substitutions (K_A/K_S ratio; see section on proteins below) than functional genes. These features can sometimes be used to recognize pseudogenes, although relatively recent pseudogenes may escape such filters.

The well-studied *Gapdh* gene and its pseudogenes illustrate the challenges¹⁵⁹. The mouse genome contains only a single functional *Gapdh* gene (on chromosome 7), but we find evidence for at least 400 pseudogenes distributed across 19 of the mouse chromosomes. Some of these are readily identified as pseudogenes, but 118 have retained enough genic structure that they appear as predicted genes in our gene catalogue. They were identified as pseudogenes only after manual inspection. The *Gapdh* pseudogenes typically have no orthologous human gene in the corresponding region of conserved synteny.

To assess the impact of pseudogenes on gene prediction, we focused on two classes of gene predictions: (1) those that lack a corresponding gene prediction in the region of conserved synteny in the human genome (2,705); and (2) those that are members of apparent local gene clusters and that lack a reciprocal best match in the human genome (5,143). A random sample of 100 such predicted genes was selected, and the predictions were manually reviewed. We estimate that about 76% of the first class and about 30% of the second class correspond to pseudogenes. Overall, this would correspond to roughly 4,000 of the predicted genes in mouse. (A similar proportion of gene predictions on chromosome 16 by Mural and colleagues⁴⁵ seem, by the same criteria, to be pseudogenes.) These two classes contain relatively few exons (average 3), and thus comprise only about 12,000 exons of the 213,562 in the mouse gene catalogue. Pseudogenes similarly arise among human gene predictions and are greatly enriched in the two classes above. This analysis shows the benefit of comparative genome analysis and suggests ways to improve gene prediction.

We also sought to identify the many additional pseudogenes that had been correctly excluded during the gene prediction process. To do so, we searched the genomic regions lying outside the predicted genes in the current catalogue for sequence with significant similarity to known proteins. We identified about 14,000 intergenic regions containing such putative pseudogenes. Most (>95%) appear to be clear pseudogenes (on the basis of such tests as ratio of non-synonymous to synonymous substitutions; see Supplementary Information and the section on proteins below), with more than half being processed pseudogenes. This is surely an underestimate of the total number of pseudogenes, owing to the limited sensitivity of the search.

Further refinement

We analysed the mouse gene predictions further, focusing on those whose best human match fell outside the region of conserved synteny and those without clear orthologues in the human genome. Two suspicious classes were identified. The first (0.4%) consists of 63 predicted genes that seem to encode Gag/Pol proteins from mouse-specific retrovirus elements. The second (about 2.5%) consists of 591 predicted genes for which the only supporting evidence

comes from a single collection of mouse cDNAs (the initial RIKEN cDNAs⁴¹). These cDNAs are very short on average, with few exons (median 2) and small ORFs (average length of 85 amino acids); whereas some of these may be true genes, most seem unlikely to reflect true protein-coding genes, although they may correspond to RNA genes or other kinds of transcripts. Both groups were omitted in the comparative analysis below.

Comparison of mouse and human gene sets

We then sought to assess the extent of correspondence between the mouse and human gene sets. Approximately 99% of mouse genes have a homologue in the human genome. For 96% the homologue lies within a similar conserved syntenic interval in the human genome. For 80% of mouse genes, the best match in the human genome in turn has its best match against that same mouse gene in the conserved syntenic interval. These latter cases probably represent genes that have descended from the same common ancestral gene, termed here 1:1 orthologues.

Comprehensive identification of all orthologous gene relationships, however, is challenging. If a single ancestral gene gives rise to a gene family subsequent to the divergence of the species, the family members in each species are all orthologous to the corresponding gene or genes in the other species. Accordingly, orthology need not be a 1:1 relationship and can sometimes be difficult to discern from paralogy (see protein section below concerning lineage-specific gene family expansion).

There was no homologous predicted gene in human for less than 1% (118) of the predicted genes in mouse. In all these cases, the mouse gene prediction was supported by clear protein similarity in other organisms, but a corresponding homologue was not found in the human genome. The homologous genes may have been deleted in the human genome for these few cases, or they could represent the creation of new lineage-specific genes in the rodent lineage—this seems unlikely, because they show protein similarity to genes in other organisms. There are, however, several other possible reasons why this small set of mouse genes lack a human homologue. The gene predictions themselves or the evidence on which they are based may be incorrect. Genes that seem to be mouse-specific may correspond to human genes that are still missing owing to the incompleteness of the available human genome sequence. Alternatively, there may be true human homologues present in the available sequence, but the genes could be evolving rapidly in one or both lineages and thus be difficult to recognize. The answers should become clear as the human genome sequence is completed and other mammalian genomes are sequenced. In any case, the small number of possible mouse-specific genes demonstrates that *de novo* gene addition in the mouse lineage and gene deletion in the human lineage have not significantly altered the gene repertoire.

Mammalian gene count

To re-estimate the number of mammalian protein-coding genes, we studied the extent to which exons in the new set of mouse cDNAs sequenced by RIKEN¹³² were already represented in the set of exons contained in our initial mouse gene catalogue, which did not use this set as evidence in gene prediction. This cDNA collection is a much broader and deeper survey of mammalian cDNAs than previously available, on the basis of sampling of diverse embryonic and adult tissues¹⁵⁰. If the RIKEN cDNAs are assumed to represent a random sampling of mouse genes, the completeness of our exon catalogue can be estimated from the overlap with the RIKEN cDNAs. We recognize this assumption is not strictly valid but nonetheless is a reasonable starting point.

The initial mouse gene catalogue of 191,290 predicted exons included 79% of the exons revealed by the RIKEN set. This is an upper bound of sensitivity as some RIKEN cDNAs are probably less than full length and many tissues remain to be sampled. On the basis of the fraction of mouse exons with human counterparts, the

percentage of true exons among all predicted exons or the specificity of the initial mouse gene catalogue is estimated to be 93%. Together, these estimates suggest a count of about 225,189 exons in protein-coding genes in mouse ($191,290 \times 0.93/0.79$).

To estimate the number of genes in the genome, we used an exon-level analysis because it is less sensitive to artefacts such as fragmentation and pseudogenes among the gene predictions. One can estimate the number of genes by dividing the estimated number of exons by a good estimate of the average number of exons per gene. A typical mouse RefSeq transcript contains 8.3 coding exons per gene, and alternative splicing adds a small number of exons per gene. The estimated gene count would then be about 27,000 with 8.3 exons per gene or about 25,000 with 9 exons per gene. If the sensitivity is only 70% (rather than 79%), the exon count rises to 254,142, yielding a range of 28,000–30,500.

In the next section, we show that gene predictions that avoid many of the biases of evidence-based gene prediction result in only a modest increase in the predicted gene count (in the range of about 1,000 genes). Together, these estimates suggest that the mammalian gene count may fall at the lower end of (or perhaps below) our previous prediction of 30,000–40,000 based on the human draft sequence¹. Although small, single-exon genes may add further to the count, the total seems unlikely to greatly exceed 30,000. This lower estimate for the mammalian gene number is consistent with other recent extrapolations¹⁴¹. However, there are important caveats. It is possible that the genome contains many additional small, single-exon genes expressed at relatively low levels. Such genes would be hard to detect by our various techniques and would also decrease the average number of exons per gene used in the analysis above.

De novo gene prediction

The gene predictions above have the strength of being based on experimental evidence but the weakness of being unable to detect new exons without support from known transcripts or homology to known cDNAs or ESTs in some organism. In particular, genes that are expressed at very low levels or that are evolving very rapidly are less likely to be present in the catalogue (R. Guigó, unpublished data).

Ideally, one would like to perform *de novo* gene prediction directly from genomic sequence by recognizing statistical properties of coding regions, splice sites, introns and other gene features. Although this approach works relatively well for small genomes with a high proportion of coding sequence, it has much lower specificity when applied to mammalian genomes in which coding sequences are sparser. Even the best *de novo* gene prediction programs (such as GENSCAN¹⁴⁵) predict many apparently false-positive exons.

In principle, *de novo* gene prediction can be improved by analysing aligned sequences from two related genomes to increase the signal-to-noise ratio¹³⁵. Gene features (such as splice sites) that are conserved in both species can be given special credence, and partial gene models (such as pairs of adjacent exons) that fail to have counterparts in both species can be filtered out. Together, these techniques can increase sensitivity and specificity.

We developed three new computer programs for dual-genome *de novo* gene prediction: TWINSKAN^{160,325}, SGP2 (refs 161, 326) and SLAM¹⁶². We describe here results from the first two programs. The results of the SLAM analysis can be viewed at <http://bio.math.berkeley.edu/slam/mouse/>. To predict genes in the mouse genome, these two programs first find the highest-scoring local mouse–human alignment (if any) in the human genome. They then search for potential exonic features, modifying the probability scores for the features according to the presence and quality of these human alignments. We filtered the initial predictions of these programs, retaining only multi-exon gene predictions for which there were corresponding consecutive exons with an intron in an aligned position in both species³²⁷.

After enrichment based on the presence of introns in aligned

locations, TWINSKAN identified 145,734 exons as being part of 17,271 multi-exon genes. Most of the gene predictions (about 94%) were present in the above evidence-based gene catalogue. Conversely, about 78% of the predicted genes and about 81% of the exons in this catalogue were at least partially represented by TWINSKAN predictions. TWINSKAN predicted an extra 4,558 (3%) new exons not predicted by the evidence-based methods. SGP2 produced qualitatively similar results. The total number of predicted exons was 168,492 contained in 18,056 multi-exon genes, with 86% of the predicted genes in the evidence-based gene catalogue at least partially represented. Approximately 83% of the exons in the catalogue were detected by SGP2, which predicted an additional 9,808 (6%) new exons. There is considerable overlap between the two sets of new predicted exons, with the TWINSKAN predictions largely being a subset of the SGP2 predictions; the union of the two sets contains 11,966 new exons.

We attempted to validate a sample of 214 of the new predictions by performing PCR with reverse transcription (RT) between consecutive exons using RNA from 12 adult mouse tissues¹⁶³ and verifying resulting PCR products by direct DNA sequencing. Our sampling involved selecting gene predictions without nearby evidence-based predictions on the same strand and with an intron of at least 1 kb. The validation rate was approximately 83% for TWINSKAN and about 44% for SGP2 (which had about twice as many new exons; see above). Extrapolating from these success rates, we estimate that the entire collection would yield about 788 validated gene predictions that do not overlap with the evidence-based catalogue.

The second step of filtering *de novo* gene predictions (by requiring the presence of adjacent exons in both species) turns out to greatly increase prediction specificity. Predicted genes that were removed by this criterion had a very low validation rate. In a sample of 101 predictions that failed to meet the criteria, the validation rate was 11% for genes with strong homology to human sequence and 3% for those without. The filtering process thus removed 24-fold more apparent false positives than true positives. Extrapolating from these results, testing the entire set of such predicted genes (that is, those that fail the test of having adjacent homologous exons in the two species) would be expected to yield only about 231 additional validated predictions.

Overall, we expect that about 1,000 (788+231) of the new gene predictions would be validated by RT-PCR. This probably corresponds to a smaller number of actual new genes, because some of these may belong to the same transcription unit as an adjacent *de novo* or evidence-based prediction. Conversely, some true genes may fail to have been detected by RT-PCR owing to lack of sensitivity or tissue, or developmental stage selection³²⁷.

An example of a new gene prediction, validated by RT-PCR, is a homologue of dystrophin (Fig. 16). Dystrophin is encoded by the

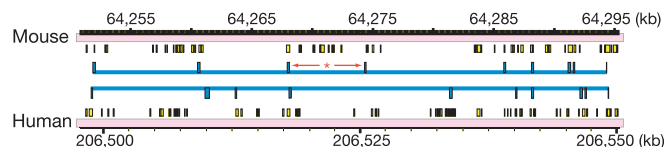


Figure 16 Structure of a new homologue of dystrophin as predicted on mouse chromosome 1 and human chromosome 2. Mouse and human gene structures are shown in blue on the chromosomes (pink). The mouse intron marked with an asterisk was verified by RT-PCR from primers complementary to the flanking exons followed by direct product sequencing³²⁷. Regions of high-scoring alignment to the entire other genome (computed before gene predictions and identification of predicted orthologues) are shown in yellow. Note the weak correspondence between predicted exons and blocks of high-scoring whole-genome alignment. Nonetheless, the predicted proteins considered in isolation show good alignment across several splice sites.

DMD gene, which is mutated in individuals with Duchenne muscular dystrophy¹⁶⁴. A gene prediction was found on mouse chromosome 1 and human chromosome 2, showing 38% amino acid identity over 36% of the dystrophin protein (the carboxy terminal portion, which interacts with the transmembrane protein β -dystroglycan). Other new gene predictions include homologues of aquaporin. These gene predictions were missed by the evidence-based methods because they were below various thresholds. These and other examples are described in a companion paper³²⁷.

The overall results of the *de novo* gene prediction are encouraging in two respects. First, the results show that *de novo* gene prediction on the basis of two genome sequences can identify (at least partly) most predicted genes in the current mammalian gene catalogues with remarkably high specificity and without any information about cDNAs, ESTs or protein homologues from other organisms. It can also identify some additional genes not detected in the evidence-based analysis. Second, the results suggest that methods that avoid some of the inherent biases of evidence-based gene prediction do not identify more than a few thousand additional predicted exons or genes. These results are thus consistent with an estimate in the vicinity of 30,000 genes, subject to the uncertainties noted above.

RNA genes

The genome also encodes many RNAs that do not encode proteins, including abundant RNAs involved in mRNA processing and translation (such as ribosomal RNAs and tRNAs), and more recently discovered RNAs involved in the regulation of gene expression and other functions (such as micro RNAs)^{165,166}. There are probably many new RNAs not yet discovered, but their computational identification has been difficult because they contain few hallmarks. Genomic comparisons have the potential to significantly increase the power of such predictions by using conservation to reveal relatively weak signals, such as those arising from RNA secondary structure¹⁶⁷. We illustrate this by showing how comparative genomics can improve the recognition of even an extremely well understood gene family, the tRNA genes.

In our initial analysis of the human genome¹, the program tRNAscan-SE¹⁶⁸ predicted 518 tRNA genes and 118 pseudogenes. A small number (about 25 of the total) were filtered out by the RepeatMasker program as being fossils of the MIR transposon, a long-dead SINE element that was derived from a tRNA^{169,170}.

The analysis of the mouse genome is much more challenging because the mouse contains an active SINE (B2) that is derived from a tRNA and thus vastly complicates the task of identifying true tRNA genes. The tRNAscan-SE program predicted 2,764 tRNA genes and 22,314 pseudogenes in mouse, but the RepeatMasker program classified 2,266 of the 'genes' and 22,136 of the 'pseudogenes' as SINEs. After eliminating these, the remaining set contained 498 putative tRNA genes. Close analysis of this set suggested that it was still contaminated with a substantial number of pseudogenes. Specifically, 19 of the putative tRNA genes violated the wobble rules that specify that only 45 distinct anticodons are expected to decode the 61 standard sense codons, plus a selenocysteine tRNA species complementary to the UGA stop codon¹⁷¹. In contrast, the initial analysis of the human genome identified only three putative tRNA genes that violated the wobble rules^{172,173}.

To improve discrimination of functional tRNA genes, we exploited comparative genomic analysis of mouse and human. True functional tRNA genes would be expected to be highly conserved. Indeed, the 498 putative mouse tRNA genes differ on average by less than 5% (four differences in about 75 bp) from their nearest human match, and nearly half are identical. In contrast, non-genic tRNA-related sequences (those labelled as pseudogenes by tRNAscan-SE or as SINEs by RepeatMasker) differ by an average of 38% and none is within 5% divergence. Notably, the 19 suspect predictions that violate the wobble rules show an average of 26%

divergence from their nearest human homologue, and none is within 5% divergence.

On the basis of these observations, we identified the set of tRNA genes having cross-species homologues with <5% sequence divergence. The set contained 335 tRNA genes in mouse and 345 in human. In both cases, the set represents all 46 expected anti-codons and exactly satisfies the expected wobble rules. The sets probably more closely represent the true complement of functional tRNA genes.

Although the excluded putative genes (163 in mouse and 167 in human) may include some true genes, it seems likely that our earlier estimate of approximately 500 tRNA genes in human is an over-estimate. The actual count in mouse and human is probably closer to 350.

We also analysed the mouse genome for other known classes of non-coding RNAs. Because many of these classes also seem to have given rise to many pseudogenes, we conservatively considered only those loci that are identical or that are highly similar to RNAs that have been published as 'true' genes. We identified a total of 446 non-coding RNA genes, which includes 121 small nucleolar RNAs, 78 micro RNAs, and 247 other non-coding RNA genes, including rRNAs, spliceosomal RNAs, and telomerase RNA. We also classified 2,030 other loci with significant similarities to known RNA genes as probable pseudogenes.

Mouse proteome

Eukaryotic protein invention appears to have occurred largely through two important mechanisms. The first is the combination of protein domains into new architectures. (Domains are compact structures serving as evolutionarily conserved functional building blocks that are often assembled in various arrangements (architectures) in different proteins¹⁷⁴.) The second is lineage-specific expansions of gene families that often accompany the emergence of lineage-specific functions and physiologies¹⁷⁵ (for example, expansions of the vertebrate immunoglobulin superfamily reflecting the invention of the immune system¹, receptor-like kinases in *A. thaliana* associated with plant-specific self-incompatibility and disease-resistance functions⁴⁹, and the trypsin-like serine protease homologues in *D. melanogaster* associated with dorsal-ventral patterning and innate immune response^{176,177}).

The availability of the human and mouse genome sequences provides an opportunity to explore issues of protein evolution that are best addressed through the study of more closely related genomes. The great similarity of the two proteomes allows extensive comparison of orthologous proteins (those that descended by speciation from a single gene in the common ancestor rather than

by intragenome duplication), permitting an assessment of the evolutionary pressures exerted on different classes of proteins. The differences between the mouse and human proteomes, primarily in gene family expansions, might reveal how physiological, anatomical and behavioural differences are reflected at the genome level.

Overall proteome comparison

We compared the largest transcript for each gene in the mouse gene catalogue to the National Center for Biotechnology Information

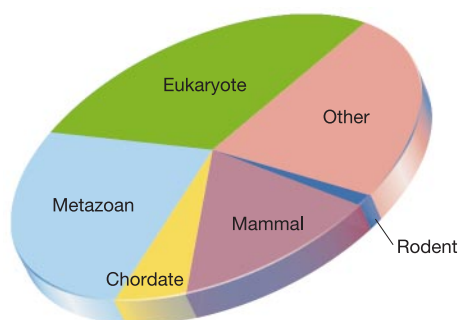


Figure 17 Taxonomic breakdown of homologues of mouse proteins according to taxonomic range. Note that only a small fraction of genes are possibly rodent-specific (<1%) as compared with those shared with other mammals (14%, not rodent-specific); shared with chordates (6%, not mammalian-specific); shared with metazoans (27%, not chordate-specific); shared with eukaryotes (29%, not metazoan-specific); and shared with prokaryotes and other organisms (23%, not eukaryotic-specific).

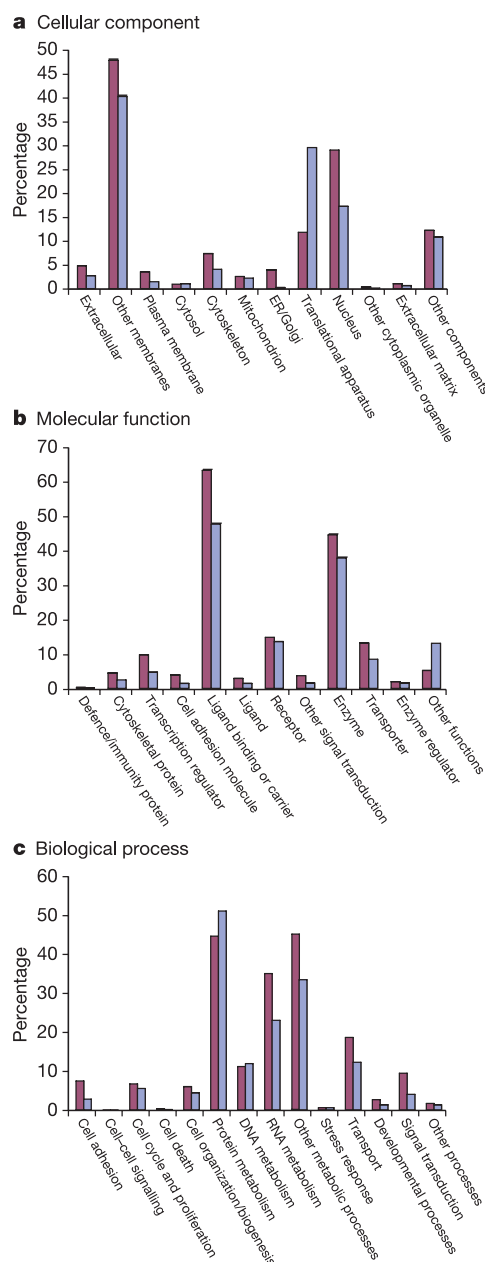


Figure 18 Gene ontology (GO) annotations for mouse and human proteins. The GO terms assigned to mouse (blue) and human (red) proteins based on sequence matches to InterPro domains are grouped into approximately a dozen categories. These categories fell within each of the larger ontologies of cellular component (a) molecular function (b) and biological process (c) (D. Hill, personal communication). In general, mouse has a similar percentage of proteins compared with human in most categories. The apparently significant difference between the number of mouse and human proteins in the translational apparatus category of the cellular component ontology may be due to ribosomal protein pseudogenes incorrectly assigned as genes in mouse.

(NCBI) database ('nr' set; ftp://ftp.ncbi.nih.gov/blast/db/nr.z) using the BLASTP program¹⁷⁸. Mouse proteins predicted to be homologues ($E < 10^{-4}$) of other proteins were classified into one of six taxonomic groupings: (1) rodent-specific; (2) mammalian-specific; (3) chordate-specific; (4) metazoan-specific; (5) eukaryote-specific; and (6) other (Fig. 17). The results were similar to those from an analysis of human proteins¹.

We annotated the current sets of mouse and human proteins with respect to the InterPro classification of domains, motifs and proteins using the InterProScan computer resource¹⁷⁹. In this way, the proteins were assigned Gene Ontology (GO) codes¹⁸⁰, which describe biological process, cellular compartment and molecular function. Comparisons of GO annotations between the two mammals showed no large-scale differences in molecular and cellular functions between the two protein sets (Fig. 18) that were not accountable by imperfections in gene prediction and annotation. Overall, about 72% of proteins contained at least one InterPro domain.

As expected, most of the protein or domain families have similar sizes in human and mouse (Table 11). However, 12 of the 50 most populous InterPro families in mouse show significant differences in numbers between the two proteomes, most notably high mobility group HMG1/2 box and ubiquitin domains. On close analysis, the differences for six of these families can be accounted for by differential expansion of endogenous retroviral sequences in the genomes. We return below to the issue of expansion of gene families.

Evolution of orthologues

To study the evolutionary forces that conserve proteins, we examined the set of 12,845 1:1 orthologues between human and mouse described above, expanding by nearly an order of magnitude the set of 1:1 orthologues used for evolutionary analysis^{14,181}. These are genes for which lineage-specific duplications seem not to have occurred in either lineage.

For each orthologous gene pair, we aligned the cDNA sequences in accordance with their pairwise amino acid alignments and

Table 11 Domain-based and family-based InterPro analysis

InterPro	Name	<i>M. musculus</i> (%)	<i>H. sapiens</i> (%)	<i>T. rubripes</i> (fish) (%)	<i>C. elegans</i> (nematode) (%)	<i>D. melanogaster</i> (insect) (%)
000276	Rhodopsin-like GPCR superfamily	3.4 (1)	2.8 (2)	1.6 (4)	2.0 (4)	0.6 (16)
000822	Zn-finger, C2H2 type	3.1 (2)	3.0 (1)	1.7 (3)	1.0 (10)	2.4 (1)
003006	Immunoglobulin/major histocompatibility complex	2.7 (3)	2.8 (3)	1.7 (2)	0.4 (32)	1.0 (6)
000719	Eukaryotic protein kinase	2.1 (4)	2.0 (4)	2.1 (1)	2.2 (2)	1.6 (3)
003593	ATPase	1.7 (5)	1.4 (5)	1.1 (5)	1.3 (8)	1.8 (2)
000504	RNA-binding region RNP-1 (RNA recognition motif)	1.4 (6)	1.0 (10)	0.7 (18)	0.6 (19)	0.9 (8)
004244	L1 transposable element	1.4 (7)	0.7 (20)	0.0 (772)	0.0 (—)	0.0 (—)
001680	G-protein beta WD-40 repeat	1.3 (8)	1.1 (7)	1.0 (7)	0.7 (16)	1.3 (5)
001841	Zn-finger, RING	1.3 (9)	1.1 (8)	0.9 (11)	0.8 (12)	0.8 (10)
000477	RNA-directed DNA polymerase (reverse transcriptase)	1.3 (10)	0.7 (19)	0.8 (14)	0.3 (42)	0.1 (163)
001849	Pleckstrin-like domain	1.2 (11)	1.0 (9)	1.0 (6)	0.4 (35)	0.5 (24)
001611	Leucine-rich repeat	1.2 (12)	0.9 (13)	0.9 (10)	0.3 (47)	0.9 (9)
001356	Homeobox	1.2 (13)	0.9 (12)	1.0 (8)	0.5 (21)	0.8 (11)
001909	KRAB box	1.1 (14)	1.1 (6)	0.0 (—)	0.0 (—)	0.0 (—)
002048	Calcium-binding EF-hand	1.1 (15)	0.9 (15)	0.8 (16)	0.4 (30)	0.7 (13)
002110	Ankyrin	1.0 (16)	1.0 (11)	0.8 (15)	0.5 (24)	0.6 (19)
001452	SH3 domain	1.0 (17)	0.8 (16)	0.8 (12)	0.3 (48)	0.5 (23)
000561	EGF-like domain	1.0 (18)	0.9 (14)	0.9 (9)	0.5 (22)	0.5 (27)
001584	Integrase, catalytic domain	0.9 (19)	0.0 (412)	0.2 (61)	0.1 (134)	0.0 (490)
003961	Fibronectin, type III	0.9 (20)	0.8 (17)	0.8 (13)	0.2 (58)	0.5 (26)
005225	Small GTP-binding protein domain	0.8 (21)	0.7 (18)	0.7 (17)	0.4 (29)	0.6 (18)
000210	BTB/POZ domain	0.8 (22)	0.6 (21)	0.6 (20)	0.8 (13)	0.5 (22)
001440	TPR repeat	0.7 (23)	0.6 (23)	0.5 (24)	0.3 (45)	0.6 (17)
001478	PDZ/DHR/GLGF domain	0.7 (24)	0.6 (22)	0.7 (19)	0.3 (43)	0.5 (25)
000008	C2 domain	0.6 (25)	0.6 (25)	0.6 (22)	0.2 (59)	0.3 (37)
000636	Cation channel, non-ligand gated	0.6 (26)	0.5 (26)	0.6 (21)	0.4 (31)	0.4 (32)
001650	Helicase, C-terminal	0.6 (27)	0.4 (31)	0.4 (32)	0.4 (27)	0.5 (21)
000980	SH2 domain	0.5 (28)	0.5 (28)	0.4 (26)	0.4 (37)	0.2 (51)
001092	Basic helix-loop-helix dimerization domain bHLH	0.5 (29)	0.4 (34)	0.4 (25)	0.2 (73)	0.4 (28)
001254	Serine protease, trypsin family	0.5 (30)	0.5 (29)	0.4 (27)	0.1 (202)	1.5 (4)
003308	Integrase, N-terminal zinc binding	0.5 (31)	0.0 (1,297)	0.0 (—)	0.0 (—)	0.0 (—)
000379	Esterase/lipase/thioesterase, active site	0.5 (32)	0.4 (33)	0.4 (30)	0.7 (15)	1.0 (7)
000626	Ubiquitin domain	0.5 (33)	0.2 (61)	0.1 (104)	0.1 (98)	0.2 (58)
004822	Histone-fold/TFIID-TAF/NF-Y domain	0.5 (34)	0.4 (30)	0.2 (82)	0.5 (26)	0.1 (155)
000387	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase	0.5 (35)	0.4 (32)	0.4 (28)	0.7 (17)	0.2 (47)
002156	RNase H	0.5 (36)	0.0 (599)	0.0 (297)	0.0 (538)	0.0 (957)
001969	Eukaryotic/viral aspartic protease, active site	0.4 (37)	0.1 (246)	0.1 (233)	0.1 (122)	0.1 (187)
001965	Zn-finger-like, PHD finger	0.4 (38)	0.3 (39)	0.3 (33)	0.2 (68)	0.3 (35)
001878	Zn-finger, CCHC type	0.4 (39)	0.2 (86)	0.2 (62)	0.2 (62)	0.2 (62)
000910	HMG1/2 (high mobility group) box	0.4 (40)	0.2 (56)	0.2 (65)	0.1 (185)	0.2 (92)
001660	Sterile alpha motif (SAM)	0.4 (41)	0.3 (47)	0.4 (31)	0.1 (166)	0.2 (52)
000483	Cysteine-rich flanking region, C-terminal	0.4 (42)	0.3 (40)	0.3 (34)	0.0 (308)	0.2 (49)
002126	Cadherin domain	0.4 (43)	0.5 (27)	0.5 (23)	0.1 (161)	0.1 (125)
000087	Collagen triple helix repeat	0.4 (44)	0.4 (36)	0.4 (29)	0.9 (11)	0.1 (132)
000372	Cysteine-rich flanking region, N-terminal	0.4 (45)	0.3 (44)	0.3 (35)	0.0 (378)	0.1 (143)
001128	Cytochrome P450	0.4 (46)	0.3 (52)	0.2 (72)	0.4 (33)	0.7 (15)
000721	Retroviral nucleocapsid protein Gag	0.4 (47)	0.0 (791)	0.0 (—)	0.0 (—)	0.0 (—)
000048	IQ calmodulin-binding region	0.4 (48)	0.4 (37)	0.3 (49)	0.1 (123)	0.2 (65)
005135	Endonuclease/exonuclease/phosphatase family	0.4 (49)	0.6 (24)	0.1 (109)	0.1 (105)	0.1 (147)
001304	C-type lectin domain	0.4 (50)	0.3 (43)	0.2 (60)	1.3 (7)	0.3 (40)

The top 50 protein families/domains in mouse are listed, and for each genome the comparative values are shown as the percentage of total genes in the genome and, in parentheses, the relative rank in the genome. This is based on InterProScan¹⁷⁹ analysis of gene products with a significant match to the InterPro collection of protein family and domain signatures. A conservative domain prediction scheme was used that excluded uncertain matches, PROSITE patterns, PRINTS predictions, and two PROSITE profiles. Only signatures of the 'family', 'domain' and 'repeat' types were considered. In the case of multiple annotated transcripts per gene, only the longest one was considered. Briefly, all InterPro hierarchical relationships among signatures with different specificity were collapsed to the broadest categories. InterPro entries in italic font represent families that have numerous members among transposons, endogenous retroviruses and pseudogenes. Significant expansions in mouse compared with human are marked in bold ($P < 0.05$ for chi-squared test on number of family/domain members with respect to total genes examined, using Dunn-Sidak corrections for multiple tests²⁸⁸). (—), indicates entries with a rank of absolute 0%.

calculated two measures of sequence evolution: the percentage of amino acid identities and the K_A/K_S ratio¹⁸². The latter quantity reflects the ratio between the rates of non-synonymous (amino-acid replacing) mutations per non-synonymous site and synonymous (silent) mutations per synonymous site (see ref. 183). Non-synonymous mutations are typically subject to strong selective pressure, whereas synonymous changes are thought typically to be neutral. Orthologue pairs generally have low values of K_A/K_S (for example, <0.05), which implies that the proteins are subject to relatively strong purifying selection¹⁸⁴. Proteins with $K_A/K_S > 1$ are formally defined as being subject to positive selection; that is, amino acid changes are accumulating faster than would be expected given the underlying silent substitution rate. However, proteins with $K_A/K_S < 1$ may still contain sites under positive selection, but the contribution of those sites to the K_A/K_S for the whole protein is offset by purifying selection at other sites¹⁸⁵. Some care is needed, however, to exclude pseudogenes in such analyses. Because pseudogenes do not encode functional proteins, the distinction between synonymous and non-synonymous mutations is irrelevant and the apparent K_A/K_S ratio will converge towards 1.

For the 12,845 pairs of mouse–human 1:1 orthologues, 70.1% of the residues were identical. The median amino acid identity was 78.5% and the median K_A/K_S ratio was 0.115 (Fig. 19 and Table 12). Most mouse and human orthologue pairs thus have a high degree of sequence identity and are under strong-to-moderate purifying selection. One consequence of the strong sequence similarity is

that computer programs such as PSI-BLAST¹⁷⁸, that use iterative alignment to detect distant homologues, gain little by using both mouse and human sequence compared with using either genome singly. For 4,344 human proteins for which no non-primate homologue could be recognized on the basis of the human sequence, the addition of a mouse orthologue added nothing new.

We sought to quantify the relative selective pressures on protein regions containing known domains. About 65% of gene pairs encode transcripts that contain at least one InterPro domain prediction (we considered only predicted domains present in corresponding positions in both orthologues). Regions containing predicted domains had higher average percentage identities and lower K_A/K_S values than regions without predicted domains or than full-length proteins (Fig. 19 and Table 12). Thus, domains are under greater purifying selection than are regions not containing domains. This is consistent with the hypothesis that domains are under greater structural and functional constraints than unstructured, domain-free regions. In this analysis (as in those below), the differences in K_A/K_S were largely due to variations in K_A (Table 12). Median K_S values clustered around 0.6 synonymous substitutions per synonymous site (Table 12), indicating that each of the sets of proteins has a similar neutral substitution rate. (These results are broadly consistent with measures of neutral substitution rate provided in the repeat and evolution sections, although the precise methodologies used and categories of sites examined affect the magnitude of estimates (see Supplementary Information).)

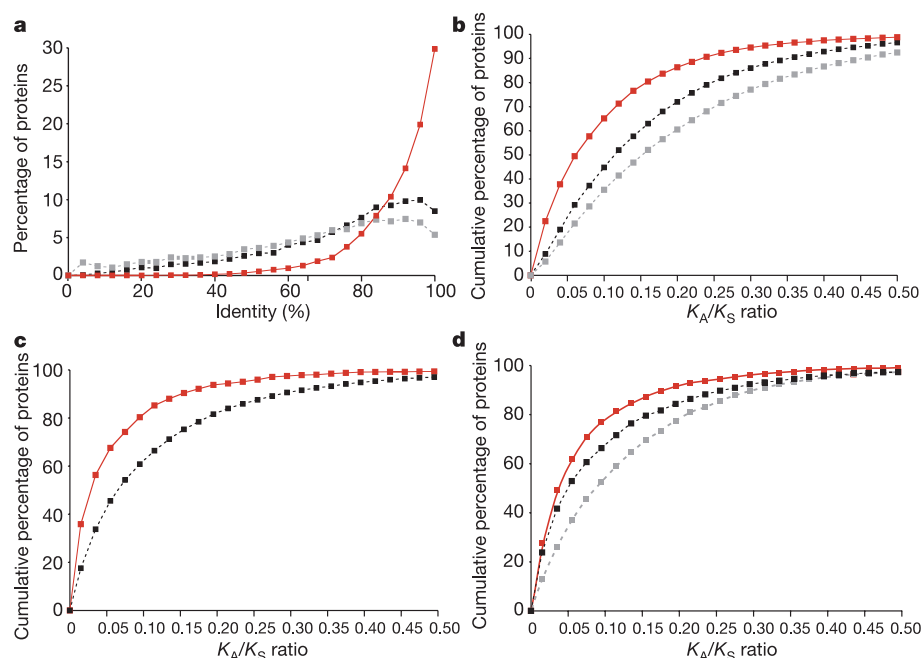


Figure 19 Protein evolution. The figure shows percentage residue identity and cumulative non-synonymous to synonymous codon rate ratios for total proteins and for regions with and without predicted InterPro domains, predicted SMART domains with or without known enzymatic activity, and SMART domains specific to three different subcellular compartments. The 12,845 orthologous gene pairs referred to in Table 12 were used for analysis. **a**, Proteins were divided into regions with and without InterPro domains, and per cent identity was calculated for total proteins (black) and for domain-containing (red line) and domain-free (grey line) regions. The higher conservation of domain-containing regions, relative to domain-free regions, is consistent with their greater functional conservation. The protein sequences are plotted in bins of 4% identity. In calculating the per cent amino acid identity between two sequences, the number of identical residues was divided by the total number of alignment positions, including positions where one sequence was aligned with a gap. **b**, Cumulative K_A/K_S ratios for total proteins (black line) and for regions with (red line) and without (grey line) predicted Interpro domains. Protein-

domain-containing regions have low K_A/K_S ratios (<0.15), suggesting that they may be subject to greater degrees of purifying selection than are the domain-free regions. The differences in functional constraints between predicted domain regions and the rest of the protein may be found to be even more pronounced, as a significant proportion of sequences may contain as yet unpredicted protein domains. **c**, Cumulative K_A/K_S ratios for SMART domain predictions with (red line) or without (black line) known enzymatic activity. The higher proportion of catalytic domains with low K_A/K_S ratios is an indication of the greater purifying selection acting on these sequences. **d**, Cumulative K_A/K_S ratios for predicted SMART domains that are specific to one of three different subcellular compartments. Compared with intracellular (cytoplasmic (red) and nuclear (black)) domains, a greater proportion of secreted domains (grey) possess higher K_A/K_S values. This indicates that secreted, often extracellular domains are subject, on average, to greater positive diversifying selection.

Table 12 K_A , K_S , K_A/K_S and pairwise percentage amino acid identities for 1:1 mouse-human orthologues

Orthologue regions	Amino acid identity (%); median (<i>n</i>); 16–83%	K_A ; median (<i>n</i>); 16–83%	K_S ; median (<i>n</i>); 16–83%	K_A/K_S ; median (<i>n</i>); 16–83%
Aligned positions in the full-length protein	78.5; (12,845); 51.0–93.1	0.071; (12,845); 0.019–0.181	0.602; (12,845); 0.414–0.981	0.115; (12,615); 0.036–0.275
Domain-containing protein regions*	93.5; (8,280); 80.6–99.1	0.032; (8,280); 0.004–0.111	0.601; (8,280); 0.390–1.072	0.061; (7,399); 0.015–0.178
Domain-free protein regions*	71.1; (12,782); 37.8–90.9	0.090; (12,615); 0.022–0.237	0.586; (12,614); 0.383–0.986	0.155; (12,035); 0.048–0.370
All predicted domains*	95.1; (17,735); 80.6–100.0	0.024; (17,735); 0–0.108	0.627; (17,735); 0.345–1.309	0.062; (13,391); 0.016–0.201
Catalytic domains†	96.6; (1,982); 86.1–100.0	0.015; (1,982); 0–0.065	0.578; (1,982); 0.346–0.979	0.033; (1,646); 0.009–0.115
Non-catalytic domains†	94.9; (15,753); 80.0–100.0	0.026; (15,753); 0–0.114	0.635; (15,753); 0.345–1.352	0.068; (11,745); 0.018–0.213
Secreted domains†	88.9; (3,901); 75.4–97.5	0.058; (3,901); 0.012–0.147	0.694; (3,901); 0.414–1.357	0.091; (3,537); 0.023–0.241
Cytoplasmic domains†	96.7; (5,795); 87.1–100	0.015; (5,795); 0–0.064	0.587; (5,795); 0.331–1.152	0.041; (4,300); 0.012–0.133
Nuclear domains†	98.6; (3,757); 85.7–100.0	0.008; (3,757); 0–0.077	0.655; (3,757); 0.302–1.696	0.050; (2,103); 0.011–0.185

*Domains predicted by Pfam and SMART.

†Domains predicted by SMART.

We next considered how the molecular functions of domains affect their evolution. Domain families with enzymatic activity were found to have a lower K_A/K_S ratio than non-enzymatic domains (Fig. 19 and Table 12). Fewer substitutions are thus tolerated in catalytic regions, suggesting that a larger proportion of amino acids contribute to substrate binding, specificity and catalysis in enzymes. Although enzymatic domains are significantly larger than non-enzymatic domains (189 compared with 47 amino acids on average), analysis indicates that there is no significant correlation between domain length and K_A/K_S ($r^2 = 0.002$).

We also examined how rates of evolution correlate with the cellular compartments in which a protein functions. We partitioned 521 of the 649 domain families in the SMART database¹⁸⁶ into secreted, cytoplasmic or nuclear classes on the basis of published data¹⁸⁷. The K_A/K_S values for the three classes showed that domains in the secreted class typically are under less purifying selection than are either nuclear or cytoplasmic domains (Fig. 19 and Table 11).

Of eight domain families with the highest (>0.15) median K_A/K_S values, six are specific to the secreted portions of proteins and are implicated in the mammalian defence and immune response system (Table 13). The fact that these proteins have the highest K_A/K_S values indicates that they are under reduced purifying selection, increased positive selection, or both. Increased positive selection may be the result of antagonistic coevolution between a mammalian host and its pathogens in a ‘genetic arms race’¹⁸⁸, where each is under strong pressure to respond to innovations in the other genome.

Mouse orthologues of human disease genes are of particular interest to biomedical research. We examined 687 human disease genes having clear orthologues in mouse¹⁸⁹. A total of 7,293 amino acid variants reported to be disease-associated¹⁹⁰ were mapped to corresponding positions in the mouse sequence. The mouse sequence was identical to the normal human sequence for 90.3%

of these positions, and it differed from both the normal and disease-associated sequence in human for 7.5% of the positions. To our surprise, the mouse sequence was identical to the human disease-associated sequence in a small number of cases (160, 2.2%). Although the causal connection with disease has not yet been proven in every one of these cases, there are at least 23 instances where the link between disease and mutation has been documented (Table 14). These include mutations in the cystic fibrosis transmembrane conductance regulator gene and the α -synuclein gene, which is associated with a familial form of Parkinson’s disease¹⁹¹. In such cases, the mouse may not provide the most appropriate model system for direct study of the mutation, although understanding the basis for the species difference may prove enlightening.

We performed a similar analysis with SNPs in coding regions of human genes. We found the location of 8,322 high-quality, coding-region SNPs from HGVbase¹⁹² within human genes using the tBLASTn computer program¹⁷⁸ and, in turn, within the corresponding positions in mouse orthologues. The mouse sequence encoded the identical amino acid as the major (more common) human allele in 67.1% of cases and as the minor human allele in 13.6% of cases. The former proportion is similar to the 70.1% of human amino acids that are conserved in mouse orthologues, indicating that most of such coding-region SNPs are not under strong selective constraint.

Evolution of gene families in mouse

As noted above, 80% of mouse proteins seem to have strict 1:1 orthologues in the human genome. Many of the remainder belong to gene families that have undergone differential expansion in at least one of the two genomes, resulting in the lack of a strict 1:1 relationship. Such gene family changes represent an insight into aspects of physiology that have emerged since the last common ancestor.

Table 13 Protein domains with high K_A/K_S values

SMART or Pfam domain family*	Domain family function	<i>n</i>	K_A/K_S median (16–83%)
CLECT	Immunity (Ly49)	126	0.150 (0.035–0.347)
IG	Immunity (immunoglobulins)	507	0.151 (0.034–0.364)
SR	Immunity (scavenger receptors)	35	0.156 (0.086–0.321)
TNFR	Immunity (CD30)	49	0.167 (0.059–0.329)
Pfam:p450	Metabolism of toxic compounds	26	0.174 (0.138–0.286)
CCP	Immunity (CD21)	100	0.181 (0.039–0.373)
SCY	Immunity (CXC chemokines)	23	0.252 (0.145–0.663)
KRAB	Transcription (ZNF133)	22	0.279 (0.051–0.468)

The eight highest K_A/K_S ratios ($K_A/K_S > 0.15$) for domain families in Pfam and SMART that are present more than 20 times in the mouse and human orthologue pairs are shown. Examples of proteins are given in parentheses.

*When equivalent families in both Pfam and SMART had median values of $K_A/K_S > 0.15$, only the SMART version is shown.

Table 14 Human disease-associated sequence variants

Disease (OMIM code)	Mutation
Hirschsprung disease (142623)	E251K
Leukencephaly with vanishing white matter (603896)	R113H
Mucopolysaccharidosis type IVA (253000)	R376Q
Breast cancer (113705)	L892S
Breast cancer (600185)	V211A, Q2421H
Parkinson's disease (601508)	A53T
Tuberous sclerosis (605284)	Q654E
Bardet-Biedl syndrome, type 6 (209900)	T57A
Mesothelioma (156240)	N93S
Long QT syndrome 5 (176261)	V109I
Cystic fibrosis (602421)	F87L, V754M
Porphyria variegata (176200)	Q127H
Non-Hodgkin's lymphoma (605027)	A25T, P183L
Severe combined immunodeficiency disease (102700)	R142Q
Limb-girdle muscular dystrophy type 2D (254110)	P30L
LCAD deficiency (201460)	Q333K
Usher syndrome type 1B (276902)	G955S
Chronic non-spherocytic haemolytic anaemia (206400)	A295V
Mantle cell lymphoma (in 208900)	N750K
Becker muscular dystrophy (300377)	H2921R
Complete androgen insensitivity syndrome (300068)	G491S
Prostate cancer (176807)	P269S, S647N
Crohn's disease (266600)	W157R

The variant amino acids of the sequence variants listed are identical in wild-type mouse orthologues

A well-documented example of family expansion is the olfactory receptor gene family, which represents a branch of the larger G-protein-coupled receptor superfamily tree^{193,194}. Duplication of olfactory receptor genes seems to have occurred frequently in both rodent and primate lineages, and differences in number and sequence have been seen as distinguishing the degrees and repertoires of odorant detection between mice and humans. Moreover, an estimated 20% of the mouse olfactory receptor homologues¹⁹⁴ and a higher percentage of human homologues^{195,196} are pseudogenes, indicating that there is a dynamic interplay between gene birth and gene death in the recent evolution of this family. The importance of these genes in reproductive behaviour is evident from

defects in pheromone responses that result from deletion of the VR1 vomeronasal olfactory receptor gene cluster¹⁹⁷.

Another example is the cytochrome P450 gene family, which is of considerable pharmacological and clinical interest. P450 cytochromes are normally terminal oxidases in multicomponent electron transfer chains, which metabolize large numbers of xenobiotic as well as endogenous compounds. Their numbers often vary among different species¹⁹⁸. This gene family is moderately but significantly expanded in mouse (84 genes) relative to human (63 genes). By comparing the cytochrome P450 gene families from mouse, human and pufferfish (*Takifugu rubripes*), we found clear expansions in four subfamilies (Cyp2b, Cyp2c, Cyp2d and Cyp4a) in mouse relative to human (Fig. 20). These occur in local gene clusters that also contain unprocessed pseudogenes. The expansions appear to be associated, in part, with gender differences in the metabolism of androgens and xenobiotics (see below).

To explore systematically recent evolution of the mouse proteome, we searched for mouse-specific gene clusters. We identified genomic regions containing four or more homologous mouse genes that descended from a single gene in the human–mouse common ancestor; these represent local expansions in the mouse lineage. To detect such clusters, we compared all transcripts of each gene with those of five genes on either side (using the BLAST-2-Sequences program with a threshold of $E < 10^{-4}$). A total of 4,563 mouse genes were found to have at least one such homologue within this window. A total of 147 such clusters containing at least four homologues was identified, of which 47 contained multiple olfactory receptor genes, which have been studied elsewhere^{193,199} and are not discussed further here. For the remaining 100 clusters, we then constructed dendrograms to examine the evolutionary relationship among the mouse proteins and their human homologues. This allowed us to identify those clusters containing mouse genes that are descendants of a single ancestral gene or for which multiple gene deletions had occurred in the human lineage.

In total, 25 such mouse-specific clusters were identified (Table 15;

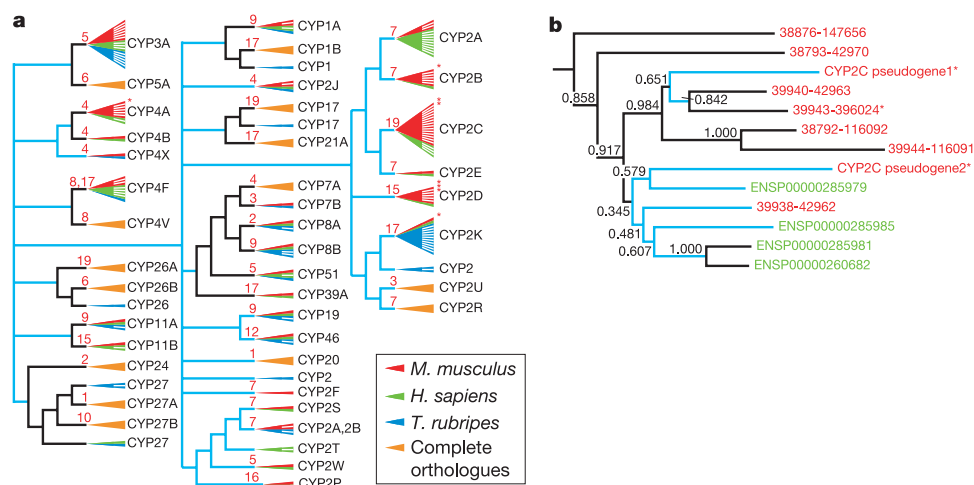


Figure 20 Cytochrome P450 protein families in mouse, human and pufferfish. **a**, Phylogenetic tree, based on the neighbour-joining method²⁹⁷, applied to the alignment of the whole P450 protein family. Each triangle represents a cytochrome P450 family cluster. Asterisks next to a triangle represent mouse pseudogenes defined by the presence of either an in-frame stop codon or a frameshift. The colour codes are indicated in the lower-right panel. When the family presents one member in each of the studied organisms, the triangle is labelled in orange. The height of the triangle is proportional to the number of proteins, which is indicated by white-line subdivisions. Chromosomal location in mouse is shown on each of the branches for each subfamily. The lengths of the branches are not drawn to scale. Branches with significant nodes (bootstrapping value >0.7) are in black, with the remainder in blue. In 6 out of the 15 CYP2C family cases, the

localization of the genomic region from which they are derived remains unassigned. **b**, Detailed phylogenetic tree of the CYP2C family based on the neighbour-joining method. The root of the tree was determined using a CYP2A sequence as out-group. Eight out of the 15 mouse CYP2C sequences are excluded in this tree as they are very short. Sequence identifiers are coloured on the basis of their source: red, mouse; green, human. Sequence identifiers followed by an asterisk indicate that the sequences contain either a premature in-frame stop codon or a frameshift. Bootstrap values are shown at the branches. Colour codes of branches are as for **a**. Although the bootstrap value for the branch containing CYP2C pseudogene2 and ENSP00000285979 is rather low (0.579), it might seem that CYP2C pseudogene2 has only recently lost its function, as a putative orthologue in human (ENSP00000285979) is still clustered with it.

Table 15 **Mouse homologous gene clusters**

Cluster	Abbreviation*	Chr†	No. of predicted genes‡	Strand (major/minor)	Function and expression	Reference
HOX cluster including Pem, Gpbox and Psx1	Hox	X	8	5/3	Probable functions in embryogenesis, placentation and rodent oogenesis. Sequence evidence for positive selection. Expression in the placenta/embryo. Pem is under androgen control in the testis and epididymis.	204, 299
Odorant binding proteins (Obp)/aphrodisin homologues (lipocalins)	Obp	X	8	4/4	Lipocalin family members probably bind volatile odorants. Aphrodisin is an aphrodisiac hormone of hamster vaginal discharges. Obps are highly expressed in the nasal area. Prostate probasin is regulated by androgens.	300, 301
Claudins	Claudins	X	6	4/2	Claudins form an intercellular barrier in tight junctions. Mutations in claudins cause reproductive defects. Four homologues were found in a testis cDNA library.	302, 303
Elafin, eppin and antileukoproteinase 1 homologues	Elafin	2	7	4/3	Reproduction-related WAP domain proteins; antimicrobial properties. Some proteins are specific to the epididymis.	304
Hydroxysteroid dehydrogenase	Hsd	3	7	7/0	Biosynthesis of hormonal steroids. Controls binding of hormone to receptor. Gender-specific expression in olfactory bulb and olfactory tubercle.	215, 216
Class mu glutathione S-transferases	GST	3	7	7/0	Conjugation of glutathione to hormones or metabolites?	305–307
Butyrophilin homologues	Butyr	4	5	3/2	Unknown functions. Contain immunoglobulin-like domain(s).	
Class Cyp4a cytochromes P450	Cyp4a	4	7	4/3	Oxidation of compounds, possibly fatty acids. Gender differences in expression (Cyp4a and Cyp4b1).	214, 308, 309
Prolactin-inducible protein; seminal vesicle-antigen (Sva)	Pip	6	4	4/0	Sva role in suppression of spermatozoa motility; 14 kDa submandibular gland protein is expressed in lacrimal, salivary and sweat glands.	
Proline-rich proteins	(–)	6	4	2/2	Salivary proteins of unknown function.	310
Submandibular gland secretory proteins	SmGSP	6	9	5/4	Salivary proteins of unknown function, related to proline-rich proteins (above). Expression is androgen-dependent.	311
Obox, family of homeobox proteins	Obox	7	6	3/3	Homeobox proteins preferentially expressed in the gonads.	312
Salivary androgen-binding protein alpha-subunit	Abpα/1 Abpα/2	7	9	7/2	Mate selection. Genes may be under positive selection due to role in subspecies recognition.	221, 222, 313, 314
Beta-defensin proteins	Bdp/1	8	5	4/1	Antimicrobial peptides. Beta-defensin 3 is expressed in salivary glands, epididymis, ovary and pancreas. Beta-defensin 5 is expressed in trachea, oesophagus and tongue.	315
Beta-defensin proteins	Bdp/2	8	5	3/2	Antimicrobial peptides. Beta-defensins 1 and 2 are expressed in kidney.	315
Carboxylesterase	CEase/1	8	6	6/0	Involved in detoxification of xenobiotics.	316
Carboxylesterase	CEase/2	8	5	4/1	Involved in detoxification of xenobiotics. Expression of egasyn is differentially regulated by androgens.	316, 317
Glioma pathogenesis-related protein homologues containing SCP domains	GPrP	10	5	3/2	Unknown.	212, 318, 319
Prolactin-related proteins	Prolactin	13	17	11/6	Placentation; probable role in development of placental blood vessels.	
Cathepsin J-like enzymes	Cath-J	13	6	6/0	Placentation; expressed in murine placenta only.	202
Eosinophil-associated ribonuclease	RNAases	14	11	7/4	Probable roles in pathogen response in eosinophils.	224, 320, 321
Class Cyp2d cytochromes P450	Cyp2d	15	5	3/2	Oxidation of compounds, possibly fatty acids. Cyp2d9 is a testosterone 16-α hydroxylase regulated by androgens. Cyp2d22 is expressed in mammary epithelial cells.	217, 218
Cystatins/steffins	Cy	16	7	5/2	Inhibitors of papain-like cysteine proteinases.	322
Proteins of unknown function	(–)	16	6	5/1	Gly-, Cys- and Tyr-rich proteins of unknown function. ESTs (BB615096 and AV261464) are derived from a testis cDNA library.	
MHC class Ib	MHCI	17	8	4/4	Some genes involved in antigen presentation, but most are not. Very different between human and mouse and between mouse strains. Class Ia peptide-binding region shows increased non-synonymous-to-synonymous substitution.	227, 323

(–), indicates that reliable alignments could not be determined, and thus they were not included in Fig. 21.

*Where paralogues were aligned into two sequence-similar subfamilies, this is indicated by appending a slash followed by the ordinal to the abbreviation.

†Mouse chromosome number.

‡Numbers of predicted genes in each cluster. In many cases this number is probably an underestimate of the true number of genes in the cluster, owing to mispredictions or incomplete genomic data, or else an overestimate owing to pseudogenes included in these totals.

see Supplementary Information). In most cases (16), the mouse-specific cluster corresponds to only a single gene in the human genome. Among these 25 clusters, two major functional themes emerge: 14 contain genes involved in rodent reproduction and 5 contain genes involved in host defence and immunity. Each of the 14 'reproduction' clusters contains at least one gene whose expression is modulated by androgens, is involved in the biosynthesis or metabolism of hormones, has an established role in the placenta, gonads or spermatozoa, or has documented roles in mate selection, including pheromone olfaction (Table 15). The fact that so many of the 25 clusters are related to reproduction is unlikely to be coincidental. Many of the most pronounced physiological differences between rodents and primates relate to reproduction, including substantial variations in placental structures, litter sizes, oestrous cycles and gestation periods. It seems likely that reproductive traits have been responsible for some of the most powerful evolutionary pressures on the mouse genome, and that the demand for innovation has been met through gene family expansions. Examination of the human genome in this way may similarly reveal gene clusters that reflect particular aspects of human reproduction.

Some of the clusters may be related to the principal differences between mice and humans in placental structure. Although both mouse and human have discoid placentae^{200,201}, they differ in the number and types of cell layers between the maternal and fetal blood. Of the expanded gene families, the cathepsin cluster on chromosome 13 and cystatins on chromosome 16 are expressed in the placenta^{202,203} and may affect its development. A non-canonical homeobox cluster on chromosome X includes *Pem*, *Psx1* and *Gpbox* (*Psx2*), which are all expressed in the placenta^{204–208}.

Another cluster is related to a different specialized aspect of reproductive physiology. This cluster, on chromosome 2, contains seminal vesicle secretory proteins that are rapidly evolving, androgen-regulated proteins involved in the formation of the copulatory plug and influence the survival and efficacy of spermatozoa^{209–211}.

Other clusters are closely related to hormone metabolism and response. These include clusters of prolactin-like genes on chromosome 13 (ref. 212), prolactin-inducible genes on chromosome 6 (refs 213, 214), 3- β -hydroxysteroid dehydrogenases on chromosome 3 (refs 215, 216), and cytochrome P450 *Cypd* genes on chromosome 15 (refs 217, 218; see Table 15).

Several of the clusters are related to olfactory cues, which have crucial roles in rodent reproduction. For example, the lipocalin-like gene cluster on chromosome X encodes proteins that are proposed to bind odorant molecules in the mucous layer overlying the receptors of the vomeronasal organ^{219,220}.

The salivary androgen-binding protein alpha (*Abp* α) pheromone gene lies within a cluster on mouse chromosome 7 that contains numerous highly related genes and pseudogenes. Males apply *Abp* α to their pelts by licking and then deposit it on their surroundings within their territory. In laboratory behavioural experiments, female mice have been shown to have a mating preference for males with a similar *Abp* α genotype, possibly to avoid interspecies breeding^{221,222}. Consequently, *Abp* α has been proposed to have a key role in the sexual isolation between *M. musculus* subspecies. The hitherto unknown *Abp* α paralogues on chromosome 7 may represent evolutionary vestiges of previously functioning *Abp* α -like molecules and/or additional functional *Abp* α -like pheromones.

Another notable cluster of probable pheromone genes was found on chromosome X. Aphrodisin is an aphrodisiac pheromone of the female hamster *Cricetus cricetus* that elicits copulatory behaviour from males²²³. The mouse chromosome X cluster contains predicted genes that are highly sequence-similar to aphrodisin and might possess similar behavioural functions.

The five mouse clusters that encode genes involved in immunity suggest that another major evolutionary force is acting on host defence genes. The five clusters include the major histocompat-

ibility complex (MHC) class Ib genes, two clusters of antimicrobial β -defensins, a cluster of WAP domain antimicrobial proteins and a cluster of type A ribonucleases. Ribonuclease A genes appear to have been under strong positive selection, possibly due to their significant role in host-defence mechanisms²²⁴. The mouse genome also contains other interesting examples of recently expanded gene clusters involved in immunity, which fall short of our strict definition of mouse-specific clusters because small families consisting of a few genes appear to have been present in the common ancestor. Examples include the *Ly6* and *Ly49* gene families, which are greatly expanded on chromosomes 15 and 6. The *Ly49* genes are of particular interest because equivalent functional niches are occupied in humans and primates by a different gene family (the non-homologous KIR family of natural killer cell receptors), an instance of convergent functional evolution^{225,226}.

The two major themes—reproduction and immunity—may not be entirely unrelated; that is, the MHC class Ib genes have roles in both pregnancy and immunity. MHC genotype is also known from ethological studies to influence mate selection, although the molecular mechanisms underlying this effect remain unknown. Within the MHC complex, the class I genes are the most divergent, having arisen after the rodent–human divergence²²⁷.

The 25 mouse-specific clusters have been generated predominantly by local gene duplication. For 74% of genes in these clusters, the most similar homologue in the mouse genome can be found either in the same cluster or within five genes from that cluster. As well as gene birth, the clusters bear witness to gene death: the *Abp* α , P450 *Cyp4a* and *Cyp4d* cytochrome P450, and carboxylesterase families all contain one or more predicted pseudogene.

Members of the clusters also seem to be undergoing rapid sequence evolution, as measured by the K_A/K_S ratio (Fig. 21). The relatively high values of K_A/K_S may reflect both positive selection (as genes diverge to take up new function) and the accumulation of mutations in moribund or dead genes. Previous studies have documented rapid evolution for a number of these clusters, including eosinophil-associated ribonucleases²²⁴, MHC class I²²⁷, class Cyp2d cytochromes P450 (ref. 228), *Abp* α subunits²²¹, the *Gpbox* homeobox cluster^{204,206} and submandibular gland secretory and proline-rich proteins²²⁹.

Genome evolution: selection

Investigation of the two principal forces that shape the evolution of the mouse and human genomes—mutation and selection—requires looking beyond coarse-scale identification of regions of conserved synteny and purely codon-based analysis of orthologues, to fine-scale alignment of the two genomes at the nucleotide level.

The substantial sequence divergence between the mouse and human genomes is still low enough that orthologous sequences undergoing neutral drift remain conserved enough for them to be aligned reliably. The challenge then is to use such alignments to tease apart the effects of neutral drift, which can teach us about underlying mutational processes, and selection, which can inform us about functionally important elements. It should be emphasized that sequence similarity alone does not imply functional constraint.

In this section, we use whole-genome alignments to explore the extent of sequence conservation in neutral sites (such as ancestral repeat sequences), known functional elements (such as coding regions) and the genome as a whole. By comparing these, we are able to estimate the proportion of regions of the mammalian genome under evolutionary selection (about 5%), which far exceeds the amount attributable to protein-coding sequences. In the next section, we then use the neutral sites to study how mutational forces vary across the genome.

Fine-scale alignment of genomes

We began by creating a catalogue of sequence alignments between the mouse and human genomes. The alignments were produced by

the BLASTZ³²⁸ program by comparing all non-repeat sequences across the genome to identify all high-scoring matches (see Supplementary Information; available for download at <http://genome.ucsc.edu/downloads.html>), then, using these as seeds, we extended the alignments into the surrounding regions, including into repeat sequences. To make the catalogue as comprehensive as possible, a given region in one genome was allowed to align to multiple, possibly non-syntentically conserved regions in the other genome. In fact, only a small proportion of the genome aligned to multiple regions (about 3.3%) or to non-syntenic regions (about 3.2%); the conclusions below are not significantly altered if we restrict attention to sequences that match uniquely in syntenic regions. Within the regions forming alignments, about 88.4% of individual human bases were aligned to bases in mouse, with the remainder aligned to indels (insertions or deletions). The alignments included approximately 98% of known coding regions, indicating that they correctly captured known, well-conserved sequence.

Regions that could be aligned clearly at the nucleotide level totalled about 1.1 Gb, corresponding to roughly 40% of the human genome (Fig. 22). This proportion may seem high if one

imagines that all such sequence conservation reflects biological function, but it does not. Simulation experiments show that DNA sequences subjected to random mutation at the neutral rate that has occurred between the human and mouse genomes (see below) can still be readily aligned by computer. In other words, most of the non-functional orthologous sequences should still be alignable. Consistent with this analysis, the alignable portion of the genomes contains a vast number of ancestral repeats, primarily relics of transposons that were present in the genome of our common ancestor with mouse and most of which are non-functional.

But if orthologous sequences should be readily alignable, the question becomes: why isn't the alignable portion much higher than 40%? In fact, the proportion is broadly consistent with what would be expected given the probable rate of turnover of sequence in the mouse and human genomes.

As a starting point, let us assume that the genome size of the last common ancestor was about 2.9 Gb (similar to the modern genomes of human and most other mammals) and let us focus only on large-scale insertions and deletions, ignoring nucleotide-level indels within aligned regions and lineage-specific duplications. These assumptions will be relaxed below.

This would imply no net change in genome size in the human lineage despite the accumulation of about 700 Mb of lineage-specific repeat sequence since the common ancestor (see section on repeats). This would require approximately 700 Mb of deletions, implying that about 24% (700 out of 2,900) of the ancestral genome was deleted and about 76% retained in the human lineage. It would also imply a net loss of about 400 Mb in the mouse lineage, despite the probable addition of about 900 Mb of lineage-specific repeat sequences, an estimate about 10% higher than that given by the

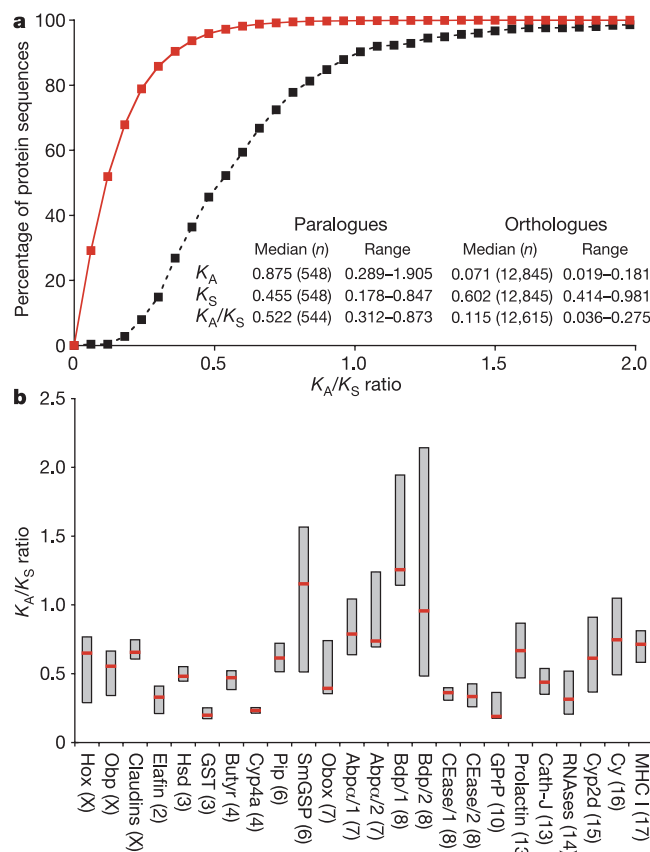


Figure 21 Protein evolution. **a**, Cumulative histogram of K_A/K_S values for locally duplicated, paralogous mouse-specific gene clusters (black boxes) in comparison with mouse-human orthologues (red boxes). The mouse-specific paralogues are more likely to be under positive diversifying selection. **b**, Box plot of K_A/K_S values for different locally duplicated, paralogous mouse-specific gene clusters. Full descriptions are found in Table 15. The chromosome on which the clusters are found is indicated in brackets after the abbreviated cluster name. The K_A/K_S values for each sequence pair in the cluster was calculated from sequences aligned using ClustalW (see Supplementary Information). The red horizontal line represents the median and the box indicates the middle 67% of the data between the 16th and 83rd percentiles. All of the paralogous clusters have median K_A/K_S values that are higher than the mouse-human orthologue median K_A/K_S (0.115), and 22 out of 25 have values greater than the 83rd percentile orthologue K_A/K_S (0.275). The Cyp2d category includes K_A/K_S values calculated separately over two sequence-similar regions in the alignment.

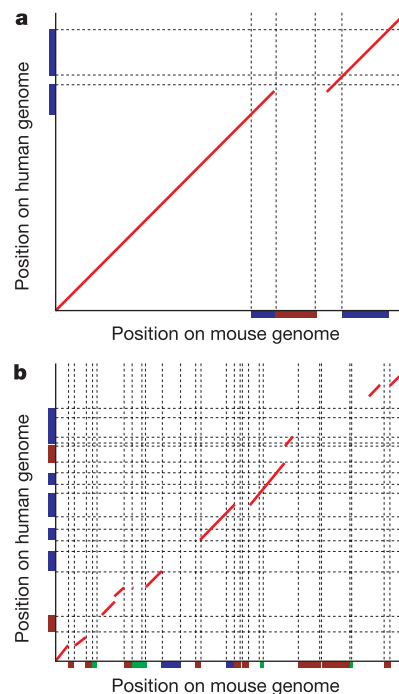


Figure 22 Dot plot showing genomic alignment between mouse and human. Typically, 40% of the human genome sequence aligns to mouse. **a**, **b**, Approximately 98% of a 2,050-bp region on human chromosome 20 aligns to the orthologous region on mouse chromosome 2 (**a**), and 56% of a 5,250-bp region on human chromosome 2 aligns to the orthologous region on mouse chromosome 1 (**b**). In both cases, the alignment skips over young/lineage-specific repeats (red boxes), but aligns through most of the ancestral repeats (blue boxes) and non-repetitive sequence (no colour). The ancestral repeats that do align are, not unexpectedly, identified as the same repeat category. Mouse also has a larger number of simple-sequence repeats (green boxes).

RepeatMasker program to allow for incomplete sensitivity in the more rapidly changing mouse genome. This would imply roughly 1,300 Mb of deletions, corresponding to the deletion of about 45% (1,330 out of 2,900) and retention of 55% of the ancestral genome.

If there was no correlation in the fixation of deletions in the two lineages, the expected proportion of the ancestral genome retained in both lineages would be about 42% ($76\% \times 55\%$). Complete independence is unlikely because deletions of functional sequences would have been selectively disadvantageous. However, deletions of modest size may largely be neutral given the relatively low proportion of functional sequence in the genome.

The estimates can be adjusted (see Supplementary Information) to account for nucleotide-level insertions and deletions and lineage-specific duplications (the expectation remains roughly the same), or to allow for different assumptions about ancestral genome size (the expectation increases by 3–4% for an intermediate size of about 2.7 Gb). This simple analysis suggests that the observed proportion of alignable genome (about 40%) is not surprising, but rather it probably reflects the actual proportion of orthologous genome remaining after the deletion in the two lineages.

In a preliminary test of this hypothesis, we identified ancestral repeats in the mouse that lay in intervals defined by orthologous landmarks. Examination of the corresponding interval in the human genome showed a rate of loss of these elements, broadly consistent with the 24% deletion rate in the human lineage assumed above (see Supplementary Information).

Such a deletion rate in the human lineage over about 75 million years is also roughly compatible with the observation that roughly 6% has been deleted over about 22 million years since the divergence from baboon, an estimate derived from the sequencing of specific regions in human and baboon (E. Green, unpublished data). Although we do not have a corresponding direct estimate of large-scale deletions in the mouse lineage, the predicted rate of about 45% is roughly twice as high as for the human lineage, which is similar to the ratio seen for nucleotide substitutions.

Rate of neutral substitution

The genome-wide alignments can be used to measure divergence rates for different types of sequence. The neutral substitution rate, for example, can be estimated from the alignment of non-functional DNA. We believe that the best representative of this class is ancestral repeat sequence, representing transposable elements inserted and fixed before the mouse–human divergence. Such ancestral repeats are more likely than any other sequence in the genome to have been under no functional constraint.

The human–mouse alignment catalogue contains approximately 165 Mb of ancestral repeat sequences, with most being clearly orthologous by alignment of adjacent non-repetitive DNA. These alignments show 66.7% sequence identity. The observed base changes can be used to infer the underlying substitution rate, which includes back mutations, by using various continuous-time Markov models²³⁰. Applying the REV model²³¹ to the ancestral repeat sites, we estimate that neutral divergence has led to between 0.46 and 0.47 substitutions per site (see Supplementary Information). Similar results are obtained for any of the other published continuous-time Markov models that distinguish between transitions and transversions (D. Haussler, unpublished data). Although the model does not assign substitutions separately to the mouse and human lineages, as discussed above in the repeat section, the roughly twofold higher mutation rate in mouse (see above) implies that the substitutions distribute as 0.31 per site (about 4×10^{-9} per year) in the mouse lineage and 0.16 (about 2×10^{-9} per year) in the human lineage.

Having established the neutral substitution rate by examining aligned ancestral repeats, we then investigated a second class of potentially neutral sites: fourfold degenerate sites in codons of genes. Fourfold degenerate sites are subject to selection in invert-

brates, such as *Drosophila*, but the situation is unclear for mammals. We examined alignments between fourfold degenerate codons in orthologous genes. The fourfold degenerate codons were defined as GCX (Ala), CCX (Pro), TCX (Ser), ACX (Thr), CGX (Arg), GGX (Gly), CTX (Leu) and GTX (Val). Thus for Leu, Ser and Arg, we used four of their six codons. Only fourfold degenerate codons in which the first two positions were identical in both species were considered, so that the encoded amino acid was identical. Slightly fewer than 2 million such sites were studied, defined in the human genome from about 9,600 human RefSeq cDNAs and aligned to their mouse orthologues. The observed sequence identity in fourfold degenerate sites was 67%, and the estimated number of substitutions per site, between 0.46 and 0.47, was similar to that in the ancestral repeat sites (see Supplementary Information).

Conservation in gene-related features

We used the genome-wide alignments to examine the extent of conservation in gene-related features, including coding regions, introns, untranslated regions, upstream regions and CpG islands.

For each type of feature, we characterized the nature of sequence conservation (including typical percentage identity, inferred substitution rates and insertion/deletion rate). We also defined a conservation score S that measures the extent to which a given window (typically 50 or 100 bp, in applications below) shows higher conservation than expected by chance. The conservation score S for an aligned region R is the normalized fraction of aligned bases that are identical (obtained by subtracting the mean and dividing by the standard deviation) and is given by:

$$S = S(R) = \frac{(p - \mu)}{\sqrt{\mu(1 - \mu)/n}}$$

where n is the number of sites within the window that are aligned, p is the fraction of aligned sites that are identical in the two genomes, and μ is the average fraction of sites that are identical in aligned ancestral repeats in the surrounding region ($\mu = 0.667$ as a genome-wide average, but, as discussed below, fluctuates locally). When the conservation score S is calculated for the set of all ancestral repeats, it has a mean of 0 (by definition) and a standard deviation of 1.19 and 1.23 for windows of 50 and 100 bp, respectively (Fig. 23). This defines the typical fluctuation in conservation score in neutral sequences. The properties of the alignments are shown in Table 16 and the distribution of conservation scores relative to neutral substitution is shown in Fig. 24.

Coding regions are distinctive in many ways. They show the highest degree of conservation (85% sequence identity or 0.165 substitutions per nucleotide site). Alignment gaps are tenfold less common than in non-coding regions. In addition, 52% of coding regions have highly significant alignments to more than one genomic region (typically, paralogues and pseudogenes), whereas only 3.3% of the genome shows such multiple alignments.

Introns are very similar, in most respects, to the genome as a whole in terms of percentage identity, gaps and multiple alignment statistics.

Conservation levels in 5' and 3' UTRs are similar to one another and intermediate between levels in coding regions and introns. The sequence identity of 75–76% is well above the intronic level of 69%. Note that our estimate of sequence identity is higher than the 70–71% reported previously¹⁸¹, in large part because that study used a global rather than a local alignment programme. The insertion and deletion characteristics of the UTRs are very similar to those of introns. Overall, 5' UTRs are slightly better conserved than 3' UTRs; however, significantly more of 3'-UTR sequence is covered by multiple alignments than 5'-UTR sequence (21% compared with 16%). This may reflect the fact that pseudogene insertion tends to proceed from the 3' end and often terminates before completion.

Promoter regions are of considerable interest. We analysed the regions located 200 bp upstream of transcription start because they

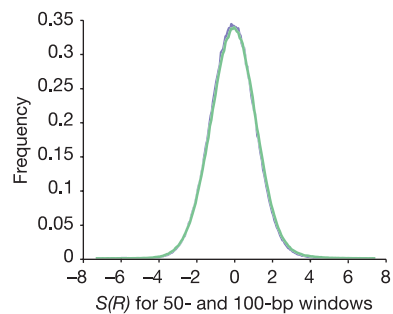


Figure 23 Distribution of the conservation score $S(R)$. The empirical distribution of $S(R)$ for all 1.9 million non-overlapping 50-bp windows (blue) containing at least 45 aligned ancestral repeat sites (standard deviation 1.19) and 1.7 million non-overlapping 100-bp windows (green) containing at least 50 aligned ancestral repeat sites (standard deviation 1.23). Both curves are bell-shaped, with a mean of zero, but the standard deviations are higher than would be expected if the sites in each window were independent and conserved with (locally estimated) probability μ . In that case the distribution of S would be approximately normal with a standard deviation of 1. Thus, these data show that there is some dependency between the substitutions within the window.

were likely to contain important promoter and regulatory signals. However, such analysis is necessarily limited by the fact that transcriptional start sites remain poorly defined for many genes. With this caveat, the upstream regions share many characteristics of 5' UTRs but have a lower percentage identity, a significantly lower proportion covered by multiple alignments, and a higher (G+C) content.

CpG islands show a conservation level similar to those of promoter and UTR regions (Fig. 24).

We also observed that levels of conservation were not uniform across these features (coding regions, introns, UTRs, upstream regions and CpG islands)²³². Figure 25 shows how conservation levels vary regionally within the features of a 'typical' gene. Sequence identity rises gradually from a background level to 78% near the approximate transcription start site, where the level reaches a plateau. It is possible that sharper definitions of transcriptional start sites would allow the footprint of the TATA box and other common structures near the transcription start site to emerge. Conversely, many human promoters lack a TATA box, and transcription start at such promoters is not typically sharply defined²³³. Sequence identity falls slowly across the 5' UTR, and then starts to rise again near the

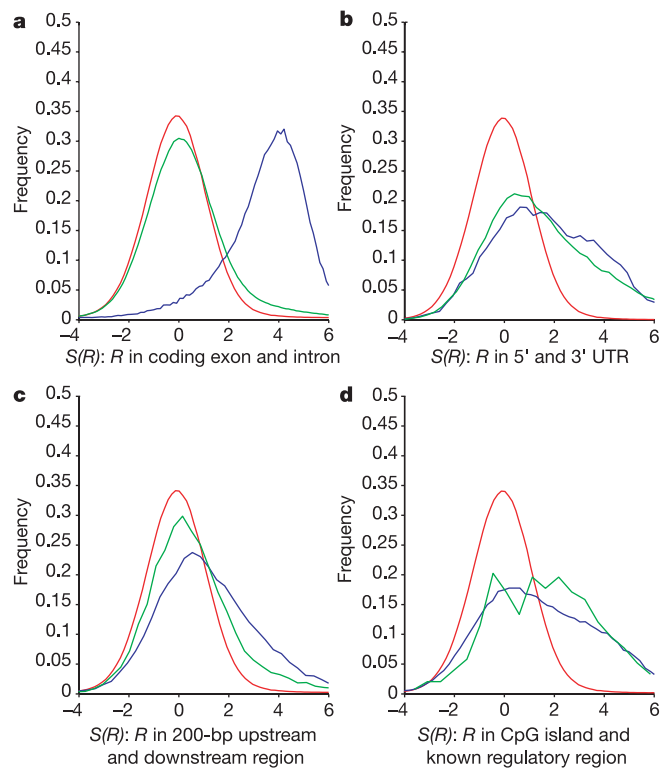


Figure 24 Comparison of histograms for conservation scores for 100-bp windows in ancient transposons (red) with 100-bp windows in other kinds of regions (blue and green). We required that at least 50 bp be aligned in each window. **a–d**, Comparisons with coding exons (blue) and introns (green) (**a**), 5' UTR (blue) and 3' UTR (green) (**b**), 200-bp upstream of transcription start (blue) and 200 bp downstream of transcription end (green) (**c**), and CpG islands (blue) and known regulatory regions (green) (**d**) are shown. The RefSeq database was used to define gene features. CpG islands were determined as discussed in the text, and known regulatory regions were collected as discussed in the text.

start codon. As expected, conservation levels rise sharply at the translation start site²³⁴, remain high throughout the coding regions, and have sharp peaks at splice sites. After the stop codon, the per cent identity is relatively low for most of the 3' UTR, but then begins to increase about 200 bases before the polyadenylation site. The

Table 16 Alignment statistics for various known features in human

Feature	Coding (%)	5' UTR (%)	3' UTR (%)	Upstream 200 bp (%)	Downstream 200 bp (%)	Intron (%)	Known regulatory regions* (%)	Ancient repeats† (%)	Genome‡ (%)
Identity (%)§	84.7	75.9	74.7	73.9	70.9	68.6	75.4	66.7	69.1
Gap initiations									
Human	0.1	0.9	1	1.1	1.2	1.1	1.2	1.1	1.1
Mouse	0.1	1.5	1.9	1.7	2.1	2	1.6	2	1.8
Gap extensions¶									
Human	0.8	4	4.6	5.3	5.8	6.5	5.5	6.4	6.2
Mouse	0.9	7.8	9.3	9.1	11.2	11.2	7	10.9	9.9
Alignment#									
X1	98.2	86.1	85.9	85.2	75	47.8	93.4	33.5	39.9
X2	52.4	16.2	20.8	11	11.4	3.5	9.7	1.2	3.3
X10	8.2	1.2	1.7	1	0.6	0.3	0	0	0.4
X100	1.5	0.3	0.4	0.1	0.2	0.04	0	0	0.1
(G+C) (%)☆	52.3	58.4	43.9	60.1	43.7	41.5	56.7	37.2	40.9

The coding, intron and UTR regions are defined by 14,729 alignments of human mRNA from the RefSeq database against the genome. Upstream 200 bp and downstream 200 bp indicate the regions 200-bp upstream and downstream of these alignments.

*From the collection of the 95 known regulatory regions described in the text.

†The ancient repeats are a collection of 2.1 million transposon relics that predate the mouse–human split, as discussed in the text.

‡The figures for the genome as a whole.

§The percentage of aligned bases in these regions that are identical.

||The number of gap initiations in the human and mouse sequences, respectively, as a percentage of the human bases in the alignments.

¶The number of gap extensions as a percentage of the human bases in the alignments.

#The percentage of human bases covered by at least 1, 2, 10 and 100 significant alignments, respectively. These numbers are taken before the last step in the construction of the alignment, when all but the best alignments for each human region are discarded.

☆The percentage of (G+C) in the human sequence.

main polyadenylation signal is AATAAA or ATTAAA positioned 10–30 bases upstream of polyadenylation²³⁵. The region of increased conservation is considerably longer than can be explained by the polyadenylation signal alone, suggesting that other 3'-UTR regulatory signals, such as those that affect mRNA stability and localization, may frequently occur near the end of the mRNA. After the polyadenylation site, there is a 30-base plateau of moderate conservation, corresponding to the weaker (T)-rich or (G+T)-rich downstream region following the polyadenylation signal.

Conservation of gene structure

We also examined the conservation of exon structure and splice signals in more detail using 1,506 pairs of human–mouse RefSeq genes confidently assigned to be orthologous (<http://www.ncbi.nlm.nih.gov/HomoloGene/>). As previously reported using smaller data sets²³⁶, overall gene structures are highly conserved between orthologous pairs: 86% of the cases (1,289 out of 1,506) have the identical number of coding exons, and 46% (692 out of 1,506) have the identical coding sequence length. When we consider all exons rather than just coding exons, we find that 941 pairs (62%) have the same number of exons. The true concordance of gene structure between the two species is probably higher, because differences will be exaggerated by differential representation of alternative splice forms between the two data sets, difficulties in mapping the cDNA sequences back to the genome, and the absence of true 5' and 3' ends.

The set of 1,289 genes with an identical number of coding exons contains 10,061 pairs of orthologous exons (plus 124 intronless genes). Exon length between orthologous exons is highly conserved: 9,131 (91%) of these human–mouse exon pairs have identical exon length. When exon pairs do have different lengths, the differences are predominantly multiples of three (858 out of the 930 with different lengths), as expected from coding-frame constraints. Nearly all orthologous exons conserve phase (10,015 or 99.5%).

In contrast, only 90 out of 8,896 orthologous introns (1%) have identical length, although there is strong correlation between the lengths of orthologous introns. Consistent with the smaller size of the mouse genome overall, orthologous mouse introns tend to be shorter. Excluding outliers, the average human intron in this data set is 4,661 bp, whereas the average mouse intron is 3,888 bp.

Within the set of 1,506 orthologous human–mouse gene pairs, there are 22 cases in which the overall coding length is identical between the gene pairs, but they differ in the number of exons. Most of these cases can be explained by a single intron insertion/deletion (Fig. 26)²³⁷, demonstrating the dynamic (but slow) evolution of gene structure.

We also found several non-canonical splice sites in the set of 8,896 orthologous introns, including RTATCCTY 5' splice signals characteristic of U12 introns, which are singularly conserved (see ref. 238 for review). We found this 5' splice signal in 20 human and 22 mouse introns from the set of 8,896, and 19 of these cases correspond to orthologous introns, indicating high levels of conservation of this distinct splicing mechanism. Also conserved are the non-canonical GC-AG introns (mechanistically identical to the GT-AG canonical introns): in the set there are 23 non-canonical GC-AG introns in human and 23 in mouse, including 19 orthologous pairs.

Conservation in known regulatory regions

We similarly sought to study the extent of conservation in regulatory control regions of genes^{232,239,240}. So far, relatively few regulatory elements have been studied extensively. We compiled a list of 95 well-characterized regulatory regions, including some liver-specific²⁴¹, muscle-specific²⁴² and general regulatory regions²⁴³. The sequences were carefully checked against the primary publications and trimmed to contain the smallest reported functional unit. The distribution of the elements was: 10% in introns, 85% in the immediate vicinity (<2 kb) of promoters, and 5% more distal from promoters. About 19% overlapped a CpG island.

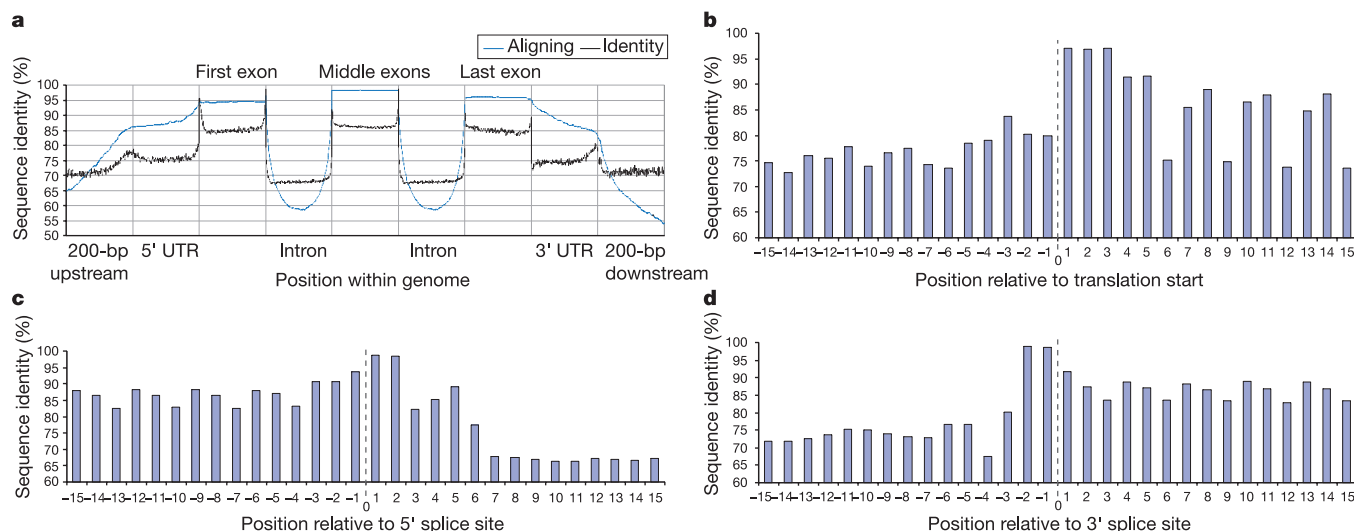


Figure 25 Variation in conservation across a gene. **a**, Conservation across a generic gene, on the basis of 3,165 human RefSeq mRNAs with known position in the genome. We sampled 200 evenly spaced bases across each of the variable-length regions labelled, resampling completely from regions shorter than 200 bp. The graph shows the average percentage of bases aligning and the average base identity when there is an alignment over each sample. There are peaks of conservation at the transition from one region to another. Here, in contrast to Table 16, only reviewed RefSeq mRNAs were used, and only those having at least 40 bases of annotated 5' and 3' UTRs. The resulting picture, however, is nearly indistinguishable from that obtained by using all RefSeq genes with at least 40 base UTRs. **b**, Conservation near translation start site using the same data set as in **a**. The bars show per cent identity of the 15 bases to either side of translation start. Note

the extreme conservation of the first codon. After this, there is substantially less conservation at the third codon position. The peak at position -3 corresponds to a purine in the Kozak consensus sequence. **c**, Conservation near the 5' splice site. The peak of conservation corresponds to the AG/GT consensus at this location, with the first G in the intron being nearly invariant. A G in the fifth base of the intron is also found in a large majority of 5' splice sites. An echo of the variation in the third codon position occurs here because it is common for exons to begin and end at codon boundaries. **d**, Conservation near the 3' splice site. Conservation in the last two bases of the intron—always AG for introns processed by the major spliceosome—is very apparent. The polypyrimidine tract beginning five bases into the intron is also visibly conserved. Once again, an echo of the variation in the third codon position can be seen.

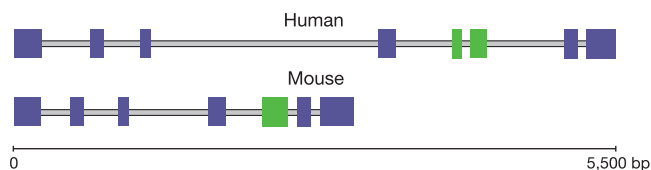


Figure 26 The human spermidine synthase gene (*SRM*) on chromosome 1, involved in the biosynthesis of polyamines, and its mouse orthologue (*Srm*) on chromosome 4. The fifth exon in the mouse gene (green) is interrupted by an intron in the human homologue. All other exons are purple.

The extent of conservation (Fig. 24 and Table 16) was considerably lower than in coding regions, but much higher than the neutral rate in ancestral repeats or than the average rate across the genome. Overall, the known regulatory regions showed a level of conservation similar to that of 5' UTRs. The (G+C) content is also substantially higher for the regulatory elements than for the genome as a whole, a property shared with exons and 5' UTRs.

Although the extent of conservation in regulatory regions—as measured by the score $S(R)$ —overlaps with that in neutral DNA (Fig. 24), this does not preclude the use of this measure to identify candidate regulatory elements. An example is given by the insulin-like growth factor binding protein acid-labile subunit gene (*IGFALS*), where the region surrounding a well-known transcription factor binding site^{244–246} stands out as unusually conserved using this measure (Fig. 27). More sophisticated models, such as Markov models on the fine texture of the alignments (matches, transitions, transversions and gaps), may discriminate regulatory regions under selection from neutrally evolving regions with better efficiency³²⁹.

Proportion of genome under selection

We then set out to investigate the fraction of a mammalian genome under evolutionary selection for biological function.

To do this, we estimated the proportion of the genome that is better conserved than would be expected given the underlying neutral rate of substitution. We compared the overall distribution

S_{genome} of conservation scores for the genome to the neutral distribution S_{neutral} of conservation scores for ancestral repeats (Fig. 23, blue curve) using a genome-wide set of 14.3 million non-overlapping 50-bp (human) windows, each containing at least 45 bp (mean 48.67 bp) of aligned sequence. The genome-wide score distribution for these windows has a prominent tail extending to the right, reflecting a substantial excess of windows with high conservation scores relative to the neutral rate (Fig. 28). The excess can be estimated by decomposing the genome-wide distribution S_{genome} as a mixture of two components: S_{neutral} and S_{selected} (reflecting windows under selection).

The mixture coefficients indicate that at least 20.8% of the windows are under selection, with the remainder consistent with neutral substitution. Because about 25.2% of all human bases are contained in the windows, this suggests that at least 5.25% (25.2% of 20.8%) of the 50-base windows in the human genome is under selection. Repeating the analysis on more stringently filtered alignments (with non-syntenic and non-reciprocal best matches removed) requiring different numbers of aligned bases per window and with 100-bp windows, yields similar estimates, ranging mostly from 4.8% to about 6.1% of windows under selection (D. Haussler, unpublished data), as does using an alternative score function that considers flanking base context effects and uses a gap penalty³³⁰. Significantly smaller window sizes, for example, 30 bp, do not provide sufficient statistical separation between the neutral and genome-wide score distributions to provide useful estimates of the share under selection.

The analysis thus suggests that about 5% of small segments (50 bp) in the human genome are under evolutionary selection for biological functions common to human and mouse. This corresponds to regions totalling about 140 Mb of human genomic DNA, although not all of the nucleotides in these windows are under selection. In addition, some bases outside these windows are likely to be under selection. In a loose sense, these regions might be regarded as containing the 'functional' conserved subset of the mammalian genome. Of course, it should be noted that non-conserved sequence may have important roles, for example, as a passive spacer or providing a function specific to one lineage. Notably, protein-coding regions of genes can account for only a fraction of the genome under selection. From our analysis of the

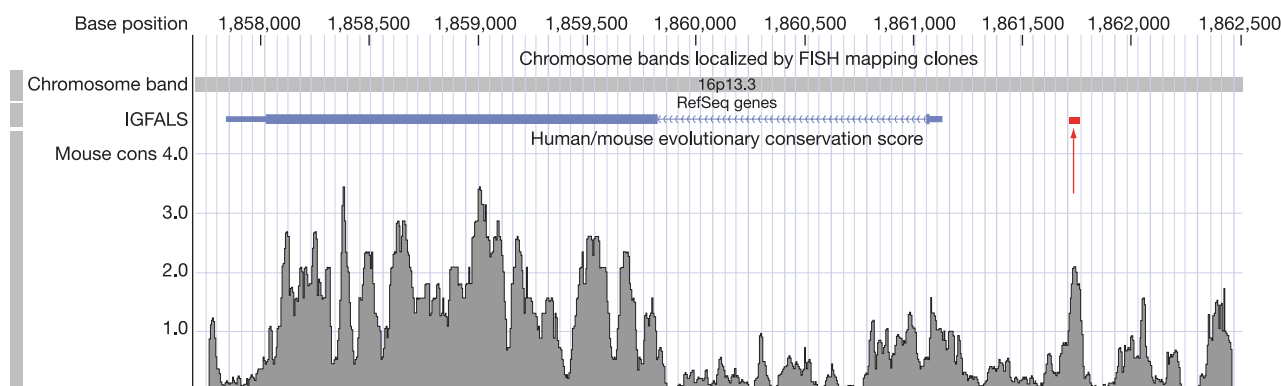


Figure 27 Conservation scores for 50-bp windows in a 4.5-kb region containing the human insulin-like growth factor binding protein acid labile subunit (*IGFALS*) gene. In the track near the top of figure, the two coding exons of the gene are displayed as taller blue rectangles, UTRs as shorter rectangles, and the intron, which separates the coding exons, is shown as a barbed line indicating direction of transcription (the gene is on the reverse strand). Log probability scores (L-scores) for all 50-bp windows are shown below the gene. The L-score is $-\log_{10}(p)$, where p is the probability under the neutral density, S_{neutral} , of getting a conservation score as high as is observed in the window. Many windows in the coding region get L-scores greater than 3, indicating less than a 1/1,000

chance of occurring under neutral evolution ($P_{\text{selected}}(S) > 0.94$; see Fig. 28), and some in a local peak in the upstream region of the gene on the right show L-scores greater than 2, indicating less than a 1/100 chance of occurring ($P_{\text{selected}}(S) > 0.75$). The red bar shows the location of the interferon- γ -activated sequence-like element (GLE), which is bound by transcription factors from the STAT5a and STAT5b protein family to control expression of this gene^{244,245}. Additional regulatory elements may be located in the other peaks of conservation. This figure is taken with permission from the UCSC browser (<http://genome.ucsc.edu>).

number and properties of genes, coding regions comprise only about 1.5% of the human genome and account for less than half of the segments under selection.

What accounts for the remainder of the genome under selection? About 1% of the genome is contained in untranslated regions of protein-coding genes, and some of this sequence is under some functional constraint. Another main class of interest are those sequences that control gene expression, such as the control element for the *IGFALS* gene shown in Fig. 27; if a typical gene contains a few such regulatory sequences, there may be tens to hundreds of thousands of such elements. In addition, conserved sequences probably encode non-protein-coding RNAs (which remain difficult to discern) and chromosomal structural elements. Furthermore, some of the conserved fraction may correspond to sequences that were under selection for some period of time but are no longer functional; these could include recent pseudogenes. In an accompanying paper, Dermitzakis and colleagues show that a large number of conserved sequences on human chromosome 21 are actively conserved but are unlikely to be genes, suggesting that a large number of non-coding sequence are under selection²⁴⁷. Characterization of the conserved sequences should be a high priority for genomics in the years ahead.

The analysis above allows us to infer the proportion of the genome under selection by decomposing the curve S_{genome} into curves S_{neutral} and S_{selected} . Importantly, it does not definitively assign an individual conserved sequence as being neutral or selected. One can calculate, for a sequence with conservation score S , the probability $P_{\text{selected}}(S)$ that the window of sequence belongs to the

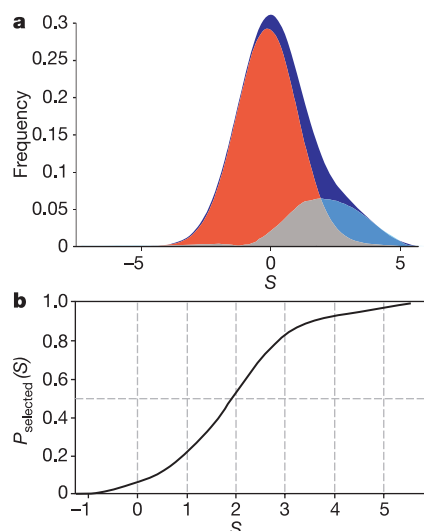


Figure 28 Proportion of the human genome under selection and the probability of a genomic window to be under selection on the basis of conservation score. **a**, The genome-wide density of conservation scores, S_{genome} (dark blue), was decomposed into a mixture of two component densities: S_{neutral} (red) and S_{selected} (light blue and grey). S_{genome} is derived from the conservation scores $S(R)$ for all windows of 50 bp in the human genome with at least 45 bases aligning to mouse. S_{neutral} is a scaled version of the S_{neutral} density from the blue curve in Fig. 23 for the 50-bp windows in ancestral repeats, representing neutrally evolving DNA. S_{selected} is the difference between the blue density and the red component, and thus represents a scaled version of S_{selected} , the predicted density for conservation scores of 50-bp windows in the human genome that are evolving under selection. The scaling factors are the estimated mixture coefficients, which are $p_0 = 0.792$ for S_{neutral} , and $1 - p_0 = 0.208$ for S_{selected} . The coefficient p_0 is calculated as the minimum of the ratio between $S_{\text{genome}}(S)$ and $S_{\text{neutral}}(S)$ for all values of S , giving a conservative estimate that maximizes the share of the mixture attributed to S_{neutral} . **b**, The probability, $P_{\text{selected}}(S)$, that a 50-bp window is under selection as a function of its conservation score $S = S(R)$. This function is derived from the mixture decomposition by setting $P_{\text{selected}}(S) = 1 - p_0 S_{\text{neutral}}(S) / S_{\text{genome}}(S)$.

selected subset (Fig. 28). The probability exceeds 83% for sequences with $S > 3$ and 93% for $S > 4$, but is only 52% for $S = 2$. In other words, some functionally important sequence cannot be separated cleanly from the tail of the distribution of neutral conservation.

How can we cleanly separate neutral and selected sequences? One solution is to extend the analysis from two species to multiple species from different branches of the mammalian radiation. Neutral sequences will tend to drift in different ways along each lineage, whereas selected sequences will tend to preserve specific sites. Multiple species comparisons should thus sharpen and separate the distributions of conservation scores, S_{neutral} and S_{selected} .

Genome evolution: mutation

Genome-wide alignments also allow us to investigate how the patterns of neutral substitution, deletion and insertion vary across the genome, providing an insight on the underlying mutational processes.

Substitution rate varies across the genome

Significant variation in the level of sequence conservation has been reported in several small-scale studies of human and mouse genomic regions^{10,248–254} and in several larger-scale studies of coding sequences^{255–260}. It has not been clear in all cases whether the variation reflects differences in neutral substitution rates or in selection. The human–mouse genome alignments allow us to address the variation more comprehensively and to test for co-variation with the rates of other processes, such as insertions of transposable elements²⁵⁵ and meiotic recombination²⁵⁸.

We used the collection of aligned ancestral repeats and aligned fourfold degenerate sites to calculate the apparent neutral substitution rate for about 2,500 overlapping 5-Mb windows across the human genome. To accurately follow fluctuations while accounting for regional changes in base composition, the regional nucleotide substitution rate in ancestral repeat sites, t_{AR} , was calculated separately for each 5-Mb window by maximum likelihood estimation of the parameters of the REV model using only the ancestral repeat sites in the window (average of about 280,000 sites per window). The regional nucleotide substitution rate in fourfold degenerate sites, t_{4D} , was calculated similarly from an average of about 3,700 fourfold degenerate sites per window. Windows with fewer than 800 ancestral repeats or fourfold degenerate sites were discarded.

The mean and standard deviations across the windows were $t_{\text{AR}} = 0.467 \pm 0.022$ and $t_{\text{4D}} = 0.447 \pm 0.067$ substitutions per site. The standard deviation is much larger (over tenfold and threefold, respectively) than would be expected from sampling variance. These data clearly indicate substantial regional fluctuation. Regional variation is also evident in comparing the average rates on different chromosomes (Fig. 29). Notably, the neutral substitution rate is lowest for chromosome X. This observation is consistent with recent reports, including our initial analysis of the human genome¹, that the mutation rate is about twofold lower in female meiosis than male meiosis. Because the proportion of time spent in the female germ line for chromosome X is 2/3 and for autosomes is 1/2, the predicted substitution rate for chromosome X should be about 8/9 or 89% of the genome-wide average. In fact, the observed ratio is 87% for fourfold degenerate sites and 92% for ancestral repeat sites. This would be consistent with (but does not prove) a roughly twofold lower mutation rate in the female germ line during the history of both the human and mouse lineages, and it explains a small amount of the variation in the genome-wide substitution rate. Nonetheless, the variability among autosomes is still much greater than could occur under a uniform substitution process, suggesting the existence of long-range factors that affect the mutation rate.

Looking at a finer scale, the two measures t_{AR} and t_{4D} are strongly

correlated across the genome (Fig. 29). They often exhibit similar behaviour across a human chromosome, as seen for human chromosome 22 (Fig. 30).

What properties of chromosomal DNA could account for the variation in substitution rate? One possible explanation is local (G+C) content, but previous studies disagree on whether it correlates strongly with divergence^{92,255,262,263}. We find that t_{AR} and t_{4D} vary with local (G+C) content, although the dependence is non-linear^{262,264} and is better fitted by regression with a quadratic curve²⁶³ (Fig. 31). In other words, the substitution rate seems to be higher in regions of extremely high or low (G+C) content, with the sign of the correlation differing in regions with high versus low (G+C) content. This pattern persists if CpG substitutions are removed from the analysis (data not shown).

Notably, t_{AR} and t_{4D} show different dependence on local (G+C) content. In particular, t_{4D} increases more sharply with high (G+C) content, whereas t_{AR} does not show as much divergence. In this and some other properties, t_{AR} and t_{4D} show differing patterns; hence they are not equivalent neutral sites. Differences in the nature of the dependence on local (G+C) content imply that the (G+C) content is a confounding variable in comparing t_{AR} and t_{4D} . Accordingly, we normalized the rates for local (G+C) content by calculating the residuals, t_{AR}^* and t_{4D}^* , with respect to the quadratic regressions above. The correspondence along chromosome 22 (a particularly (G+C)-rich chromosome) is markedly enhanced (r^2 increases from 0.55 to 0.75) by this correction (Fig. 30), as is the overall genome-wide correlation (r^2 increases from 0.22 to 0.33).

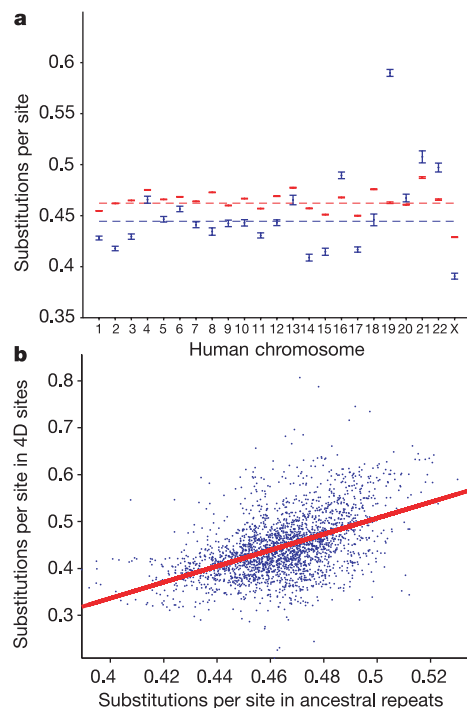


Figure 29 Estimated average number of substitutions per site in ancestral repeat sites (t_{AR}) (red) and in fourfold degenerate (4D) sites (t_{4D}) (blue) for each human chromosome. **a**, Estimates are made from the REV model using all aligned sites of the given type in the chromosome. Dashed lines show the genome-wide averages. Human chromosome 19 is a conspicuous outlier for its very large number of substitutions in fourfold degenerate sites (also noted in ref. 259); notably, its substitution rate in ancestral repeat sites is normal. Chromosome X shows lower rates of substitution in both types of sites, consistent with the observation that the male mutation rate is approximately twice the female rate¹ (see text). Variability in neutral rates among autosomes is significant, as noted in ref. 13. **b**, Scatter plot of t_{AR} against t_{4D} for 2,424 5-Mb windows in the human genome with at least 800 aligning sites. The red line is the linear regression line ($r^2 = 0.22$; $P < 10^{-6}$).

Substitution rate co-varies with other evolutionary rates

In addition to nucleotide substitutions, genomes evolve by insertion (primarily of transposable elements) and deletion. We examined the rate of deletion in the mouse genome, as measured by the fraction of non-aligning ancestral human DNA (NA_{anc}). Although some of the non-alignable sequence may represent lineage-specific insertions not detected by RepeatMasker (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>)¹⁷⁷ or failure to align some orthologous sequences, the great bulk probably represents deletions in the mouse genome. The fraction NA_{anc} varies markedly across overlapping windows of 5 Mb, with a range from 0.295 to 0.985 and mean and standard deviation 0.521 ± 0.095 .

We also examined the rate of insertion (and retention) in the human genome since its divergence from mouse, as measured by the proportion of lineage-specific repeats in overlapping 5-Mb windows across the human genome. The overall level of insertion and retention showed substantial variation across the genome, ranging from 0.159 to 0.805 with a mean of 0.290 ± 0.063 . To avoid complications from the tendency of some repeats, such as Alus, to be selectively removed from some regions of the genome¹, we used one family of repeats, the LTRs, to monitor the relative frequency of insertion and retention. Similar to repeats as a whole, the fraction of

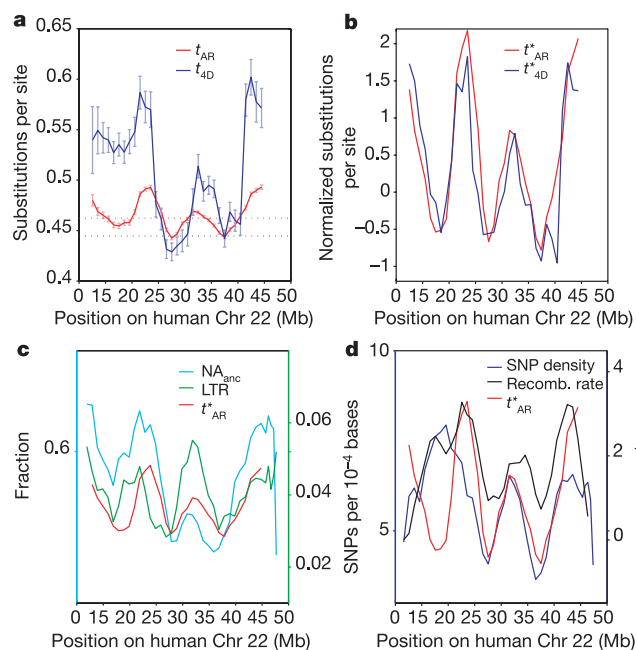


Figure 30 Variation in features along human chromosome 22. **a**, Variation in t_{AR} (red) and t_{4D} (blue) in 5-Mb windows, overlapping by 4-Mb, along human chromosome 22. Only windows with at least 800 aligned fourfold degenerate sites and 800 aligned ancestral repeat sites are shown. The position of the window is plotted at the midpoint. Horizontal dotted lines indicate the genome-wide estimates of t_{AR} and t_{4D} . Confidence intervals were computed on the basis of the number of ancestral repeat and fourfold degenerate sites aligning in each window; points where the confidence interval does not overlap the genome-wide estimate indicate windows with significant differences in evolutionary rate. **b**, Similar to **a**, but with t_{AR}^* and t_{4D}^* , the normalized rates obtained taking residuals of t_{AR} and t_{4D} from the quadratic functions of (G+C) content shown in Fig. 31. **c**, Fraction of DNA (blue) that is not in lineage-specific repeats identified by RepeatMasker and does not align to mouse, NA_{anc} , and the fraction of DNA (green) contained in human lineage-specific LTR repeats identified by RepeatMasker, along with t_{AR}^* (red), calculated in overlapping 5-Mb windows as in **b**. **d**, SNP density (blue) in each overlapping 5-Mb window (average number of SNPs per 10 kb) calculated using SNPs from random reads (The SNP Consortium website; data were collected in July 2002, <http://snp.cshl.org>). The average recombination rate (black) in each 5-Mb window, in cM per Mb, estimated from the deCode genetic map²⁶⁹ is shown, as well as t_{AR}^* (red), calculated in overlapping 5-Mb windows as in **b**.

each window occupied by lineage-specific LTRs varies substantially across the human genome, ranging from 0 to 0.378, with a mean of 0.0598 ± 0.0197 .

All three forces that alter the genome (nucleotide substitution, deletion and insertion) thus vary substantially across the genome. Moreover, they are significantly correlated and tend to co-vary along chromosomes (Fig. 30 and Table 17). Notably, these three measures of interspecies divergence are also correlated with recent substitutions in the human genome, as measured by the density of SNPs identified by the SNP Consortium²⁶⁵ (Fig. 30).

Furthermore, recent studies report that divergence at fourfold degenerate sites and SNP frequency are both correlated with the local rate of meiotic recombination^{258,266–268}. We examined the relationship between our measures of genome-wide divergence and recombination rate using recently reported high-resolution measurements of recombination rates in the human genome²⁶⁹. Both measures of neutral substitution rate and SNP rate showed a significant correlation with recombination rate (Fig. 30 and Table 17).

The correlations above are not explained by co-variation with local (G+C) content. All except the correlation between SNP frequency and LTR insertion rate remain significant when dependence on underlying human (G+C) content is factored out by taking the residuals of a quadratic regression on regional human (G+C) content; indeed, the correlations are for the most part enhanced (Table 17). Similarly, correlations remain significant when the difference between the (G+C) content of orthologous mouse and human regions is also factored out²⁶¹. Thus, (G+C) content changes between mouse and human, as explored previously²⁵⁹, do not adequately explain the correlations. Finally, to

obtain more rigorous estimates of significance, the correlations were re-evaluated on non-overlapping sets of 5-Mb windows, and on non-overlapping 1-Mb windows as well, with similar results²⁶¹.

Possible explanations for variation

What explains the correlation among these many measures of genome divergence? It seems unlikely that direct selection would account for variation and co-variation at such large scales (about 5 Mb) and involving abundant neutral sites taken from ancestral transposon relics. Selection against deleterious mutations can remove linked polymorphisms^{270,271}, but it is not clear that such effects or related effects²⁷² could extend to such large scales or to interspecies divergence over such large time periods²⁷³.

It seems more probable that these features reflect local variation in underlying mutation rate, caused by differences in DNA metabolism or chromosome physiology. The causative factors may include recombination-associated mutagenesis^{258,266}, transcription-associated mutagenesis²⁷⁴, transposon-associated deletion and genomic rearrangement^{275–278}, and replication timing^{279,280}. Nuclear location may also be involved, including proximity to matrix attachment sites, heterochromatin, nuclear membrane, and origins of replication.

It is clear that the mammalian genome is evolving under the influence of non-uniform local forces. It remains an important challenge to unravel the mechanistic basis and evolutionary consequences of such variation.

Genetic variation among strains

To facilitate genetic mapping studies, it would be valuable to create a mouse genetic map based on SNPs. The use of SNPs would allow the generation of an even denser map, which would allow mouse geneticists to fully exploit the recombinational resolution that can be achieved in large crosses. A cross with 2,000 meioses divides the genome (with a genetic length of about 16 morgans) into approximately 32,000 distinct recombinational 'bins' and it would be convenient to have an even higher density of genetic markers available for fine-scale mapping. In addition, SNPs offer potential advantages in terms of automation and parallelism^{265,281,282}.

Given a reference sequence of the B6 strain, it is straightforward to find SNPs relative to any other strain. One simply needs to generate random shotgun reads from the strain, align them to the reference sequence and search for high-quality sequence differences.

As a pilot project, we created initial SNP collections from three strains: 129S1/SvImJ (129), C3H/HeJ (C3H) and BALB/cByJ (BALB) (Table 18). So far we have identified 47,279 high-quality candidate SNPs between the 129 and B6 strains, 20,294 SNPs between C3H and B6 and 11,696 between BALB and B6. The initial SNP collection thus contains more than 79,000 SNPs. This total is expected to grow with deeper coverage and the inclusion of

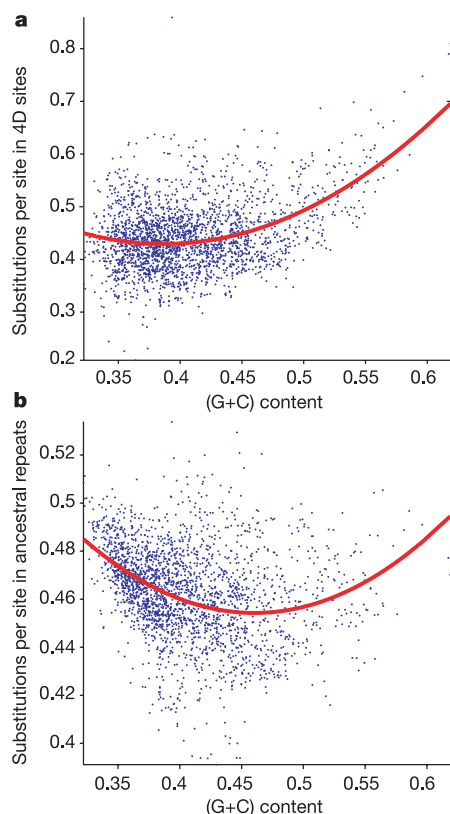


Figure 31 Expected number of substitutions in fourfold degenerate (4D) sites (**a**) and ancestral repeat sites (**b**) plotted against human (G+C) content. The second-order (quadratic) polynomial regression curve is shown in red.

Table 17 Pairwise correlations for six divergence features

	Neutral sub (AR)	Neutral sub (4D)	Deletion	SNP	Recombination
Original variables					
Neutral sub (4D)	0.49	—	—	—	—
Deletion	0.426	0.482	—	—	—
SNP	0.524	0.2	0.073	—	—
Recombination	0.222	0.24	0.031	0.244	—
Insertion	0.267	0.042	0.409	0.058	−0.228
Residuals from quadratic regression on (G+C)					
Neutral sub (4D)	0.584	—	—	—	—
Deletion	0.446	0.394	—	—	—
SNP	0.533	0.305	0.12	—	—
Recombination	0.364	0.184	−0.008	0.322	—
Insertion	0.157	0.167	0.521	−0.008	−0.091

AR, ancestral repeat; 4D, fourfold degenerate site.

Table 18 SNPs generated by WGS sequencing

Strain	Reads*	SNPs
129S1/SvImJ	67,974	47,279
C3H/HeJ	34,949	20,294
BALB/cByJ	19,686	11,696
Total	122,609	79,269

* Reads passing all filters: sequence quality, SSAHA-SNP program³⁰⁴ and unique placement on the genome.

additional strains. We tested a random sample of 83 candidate SNPs by resequencing and found that all 83 were authentic, indicating that most of the candidate SNPs are true variants.

The average density of SNPs between B6 and each of the three strains was in the range 1 per 500–700 bp. The distribution of SNPs is highly non-uniform (consistent with earlier observations²⁸²). Some regions of the genome appear to be unusually rich in SNPs, whereas others are devoid of SNPs.

In an accompanying paper, Wade and colleagues²⁸³ analyse this non-uniform distribution of SNPs and demonstrate that genetic variation between strains occurs in a harlequin pattern of alternating blocks of either high or low SNP rate, typically extending more than 1 Mb. Genotyping of additional strains reveals that the SNPs largely represent alternative alleles from *M. m. domesticus* and *M. m. musculus*, and that the blocks probably represent the distinct segmental contributions of the two subspecies to existing laboratory mouse strains. Detailed knowledge of these blocks can thus allow reconstruction of the history and relationship among mouse strains. Furthermore, it can be used to perform association studies on mouse strains, by correlating differences in phenotype across multiple strains with the underlying block structure of genetic variation.

Implications for the laboratory mouse

The promise of genomics is the ability to connect phenotypes with genotypes for a wide variety of traits and to use the resulting molecular insights to develop new approaches for the cure and prevention of disease. The laboratory mouse occupies a central place in this vision, both as a prototype for all mammalian biology and as a well-characterized organism for modelling human disease states^{15,16,123}. In this section, we briefly discuss ways in which the mouse genome sequence will accelerate biomedical progress in the future. Because the sequence has been made available in public databases in advance of publication, examples for many of the predictions can already be cited.

Positional cloning of genes for mendelian phenotypes

More than 1,000 spontaneously arising and radiation-induced mouse mutants causing heritable mendelian phenotypes are catalogued in the Mouse Genome Informatics (MGI) database (<http://www.informatics.jax.org>). Largely through positional cloning, the molecular defect is now known for about 200 of these mutants. The availability of an annotated mouse genome sequence now provides the most efficient tool yet in the gene hunter's toolkit. One can move directly from genetic mapping to identification of candidate genes, and the experimental process is reduced to PCR amplification and sequencing of exons and other conserved elements in the candidate interval. With this streamlined protocol, it is anticipated that many decades-old mouse mutants will be understood precisely at the DNA level in the near future. An example of how the draft genome sequence has already been successfully used is the recent identification of the mouse mutation 'chocolate' in the melanosome protein Rab38 (ref. 284).

The mouse genome sequence will be even more crucial in efforts to exploit the growing repertoire of mutant mice being generated by chemical mutagenesis with *N*-ethyl-*N*-nitrosurea (ENU) and other agents. At least ten large-scale ENU mutagenesis centres have recently been established worldwide, focusing on dominant or

recessive screens for a wide variety of viable, clinically relevant phenotypes¹⁵. Hundreds of new mutants with biochemical, development and behavioural phenotypes are being generated each year. For each mutant, identification of the molecular cause will require positional cloning.

Another means of generating mutants, the so-called 'gene trap' approach, uses a promoterless reporter construct for random insertion into the genome of embryonic stem cells. Expression of the reporter correlates with integration into a transcriptional unit, which is disrupted by the event and confers its tissue and developmental specificity to the reporter. Several large-scale gene-trap programmes are underway worldwide¹⁵. Availability of the genome sequence now makes the determination of the precise integration site in an interesting mutant an almost trivial exercise.

Identification of quantitative trait loci

The availability of more than 50 commonly used laboratory inbred strains of mice, each with its own phenotype for multiple continuously variable traits, has provided an important opportunity to map QTLs that underlie heritable phenotypic variation. A systematic initiative is currently underway²⁸⁵ to define parameters such as body weight, behavioural patterns, and disease susceptibility among a standard set of inbred lines, and to make these data freely available to the scientific community in the Mouse Phenome Database (www.jax.org/phenome). Appropriate crosses between such lines, followed by genotyping, will enable the mapping of QTLs, which can then be subjected to positional cloning. The degree of difficulty is substantially greater for a QTL cloning project than for a mendelian disorder, however, as the responsible intervals are usually much larger, the boundaries more difficult to delineate precisely, and the causative variant often much more subtle²⁸⁶. For these reasons, only a handful of the approximately 1,000 mapped QTLs have been identified at the molecular level. The availability of the mouse sequence should greatly improve the chances for future success.

Success in QTL identification will be enhanced if genetic mapping can be combined with genomic sequence, expression array data and proteomic data. Furthermore, the use of high-density SNP maps to identify blocks of ancestral identity among mouse strains and to correlate them with phenotypes may assist in the design of QTL experiments. The availability of BAC libraries from several strains will facilitate testing candidate genes for QTLs through the construction of transgenic mice²⁸⁷. The combination of multiple perspectives on genome sequence, variation and function should thus provide a powerful platform for revealing molecular mechanisms of phenotypic variation.

Creation of knockout and knockin mice

The wide application of homologous recombination in embryonic stem cells has provided a remarkable abundance of 'custom' mice with specifically engineered loss- or gain-of-function mutations in specific genes of biological or medical interest. Yet this remains a time-consuming process. The design of recombinant DNA constructs for injection has often been delayed by incomplete knowledge of gene structure, requiring tedious restriction mapping or sequencing, and occasionally giving rise to unsatisfying outcomes due to incorrect information. The availability of the mouse genome sequence will both speed the design of such constructs and reduce the likelihood of unfortunate choices. Furthermore, the long-range continuity of the sequence should facilitate the generation of models of contiguous gene-deletion syndromes.

Creation of transgenic animals

For many transgenic experiments, it is important to maintain copy-dependent, tissue-specific expression of the transgene. This is most readily accomplished through BAC transgenesis. The availability of a deep, end-sequenced BAC library from the B6 strain mapped to

the genome sequence now makes it straightforward to obtain a desired gene in a BAC for such experiments; end-sequenced BAC libraries from other strains should be available in the future. BACs also provide the ability to make mutant alleles with relative ease, by taking advantage of powerful genetic engineering techniques for custom mutagenesis in the *Escherichia coli* host.

Applications to cancer

The mouse genome sequence also has powerful applications to the molecular characterization of the somatic mutations that result in neoplasia. High-density SNP mapping to identify loss of heterozygosity^{288,289}, combined with comparative genomic hybridization using cDNA or BAC arrays^{290,291}, can be used to identify chromosomal segments showing loss or gain of copy number in particular tumour types. The combination of such approaches with expression arrays that include all mouse genes should further enhance the ability to pinpoint the molecular lesions that result in carcinogenesis. Full sequencing of all the exons and regulatory regions of known tumour suppressors, oncogenes, and other candidate genes can now be contemplated, as has been initiated in a few centres for human tumours²⁹².

As a specific example of the use of the draft sequence for oncogene discovery, several groups recently used retroviral infection in mice to recover new cancer susceptibility loci. The ability to compare rapidly retrieved sequence tags to the draft genome sequence greatly accelerated the process of cancer gene discovery^{293–295}.

Making better mouse models

Not all mouse models replicate the human phenotype in the expected way. The availability of the full human and mouse sequences provides an opportunity to anticipate these differences, and perhaps to compensate for them. In some instances, it may turn out that the murine mutation did not reside in the true orthologue of the human disease gene. Alternatively, in a circumstance where the human genome contains only a single gene family member, but the mouse genome contains a paralogue as well as the orthologue, one can anticipate that knockout of the orthologue alone may give a much milder phenotype (or none at all). Such was the case, for instance, with the oculocerebrorenal syndrome described by Lowe and colleagues²⁹⁶. Creating double knockout mice may then provide a closer match to the human disease phenotype.

Understanding gene regulation

Of the approximately 5% of windows of the mammalian genome that are under selection, most do not appear to code for protein. Much of this sequence is probably involved in the regulation of gene expression. It should be possible to pinpoint these regulatory elements more precisely with the availability of additional related genomes. However, mouse is likely to provide the most powerful experimental platform for generating and testing hypotheses about their function. An example is the recent demonstration, based on mouse–human sequence alignment followed by knockout manipulation, of several long-range locus control regions that affect expression of the *Il4/Il13/Il5* cluster⁴.

Conclusion

The mouse provides a unique lens through which we can view ourselves. As the leading mammalian system for genetic research over the past century, it has provided a model for human physiology and disease, leading to major discoveries in such fields as immunology and metabolism. With the availability of the mouse genome sequence, it now provides a model and informs the study of our genome as well.

Comparative genome analysis is perhaps the most powerful tool for understanding biological function. Its power lies in the fact that evolution's crucible is a far more sensitive instrument than any other available to modern experimental science: a functional alteration

that diminishes a mammal's fitness by one part in 10⁴ is undetectable at the laboratory bench, but is lethal from the standpoint of evolution.

Comparative analysis of genomes should thus make it possible to discern, by virtue of evolutionary conservation, biological features that would otherwise escape our notice. In this way, it will play a crucial role in our understanding of the human genome and thereby help lay the foundation for biomedicine in the twenty-first century.

The initial sequence of the mouse genome reported here is merely a first step in this intellectual programme. The sequencing of many additional mammalian and other vertebrate genomes will be needed to extract the full information hidden within our chromosomes. Moreover, as we begin to understand the common elements shared among species, it may also become possible to approach the even harder challenge of identifying and understanding the functional differences that make each species unique. □

Methods

Production of sequence reads

Paired-end reads from libraries with different insert sizes were produced as previously described¹ using 384-well trays to ensure linkages.

Availability of sequence and assembly data

Unprocessed sequence reads are available from the NCBI trace archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/mus_musculus/). Raw assembly data (before removal of contaminants, anchoring to chromosomes, and addition of finished sequence) are available from the Whitehead Institute for Biomedical Research (WIBR) (ftp://wolfram.wi.mit.edu/pub/mouse_contigs/Mar10_02/). The released assembly MGSCv3 is available from Ensembl (http://www.ensembl.org/Mus_musculus/), NCBI (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/MGSCv3_Release1/), UCSC (<http://genome.ucsc.edu/downloads.html>) and WIBR (ftp://wolfram.wi.mit.edu/pub/mouse_contigs/MGSC_V3/). (See Supplementary Information for detailed Methods.)

Received 18 September; accepted 31 October 2002; doi:10.1038/nature01262.

1. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. O'Brien, S. J. *et al.* The promise of comparative genomics in mammals. *Science* **286**, 458–462, 479–481 (1999).
4. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
5. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
6. Oeltjen, J. C. *et al.* Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7**, 315–329 (1997).
7. Ellsworth, R. E. *et al.* Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc. Natl Acad. Sci. USA* **97**, 1172–1177 (2000).
8. Mallon, A. M. *et al.* Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res.* **10**, 758–775 (2000).
9. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**, 104–111 (2001).
10. DeSilva, U. *et al.* Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.* **12**, 3–15 (2002).
11. Toyoda, A. *et al.* Comparative genomic sequence analysis of the human chromosome 21 down syndrome critical region. *Genome Res.* **12**, 1323–1332 (2002).
12. Ansari-Lari, M. A. *et al.* Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* **8**, 29–40 (1998).
13. Lercher, M. J., Williams, E. J. & Hurst, L. D. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**, 2032–2039 (2001).
14. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
15. Rossant, J. & McKerlie, C. Mouse-based phenogenomics for modelling human disease. *Trends Mol. Med.* **7**, 502–507 (2001).
16. Paigen, K. A miracle enough: the power of mice. *Nature Med.* **1**, 215–220 (1995).
17. Hogan, B., Beddington, R., Costantini, F. & Lacy, E. *Manipulating the Mouse Embryo: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Woodbury, New York, 1994).
18. Joyner, A. L. *Gene Targeting: A Practical Approach* (Oxford Univ. Press, New York, 1999).
19. Copeland, N. G., Jenkins, N. A. & Court, D. L. Recombineering: a powerful new tool for mouse functional genomics. *Nature Rev. Genet.* **2**, 769–779 (2001).
20. Yu, Y. & Bradley, A. Engineering chromosomal rearrangements in mice. *Nature Rev. Genet.* **2**, 780–790 (2001).
21. Bucan, M. & Abel, T. The mouse: genetics meets behaviour. *Nature Rev. Genet.* **3**, 114–123 (2002).

22. Silver, L. M. *Mouse Genetics: Concepts and Practice* (Oxford Univ. Press, New York, 1995).
23. Bromham, L., Phillips, M. J. & Penny, D. Growing up with dinosaurs: molecular dates and the mammalian radiation. *Trends Ecol. Evol.* **14**, 113–118 (1999).
24. Nei, M., Xu, P. & Glazko, G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA* **98**, 2497–2502 (2001).
25. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
26. Madsen, O. *et al.* Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610–614 (2001).
27. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).
28. Keeler, C. E. *The Laboratory Mouse: Its Origin, Heredity and Culture* (Harvard Univ. Press, Cambridge, Massachusetts, 1931).
29. Morse, H. *The Mouse in Biomedical Research* (eds Foster, H. L., Small, J. D. & Fox, J. G.) 1–16 (Academic, New York, 1981).
30. Morse, H. C. *Origins of Inbred Mice* (ed. Morse, H. C.) 1–21 (Academic, New York, 1978).
31. Haldane, J. B. S., Sprunt, A. D. & Haldane, N. M. Reduplication in mice. *J. Genet.* **5**, 133–135 (1915).
32. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
33. Dietrich, W. *et al.* *Genetic Maps* (ed. O'Brien, S.) 4.110–4.142, (1992).
34. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
35. Love, J. M., Knight, A. M., McAleer, M. A. & Todd, J. A. Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. *Nucleic Acids Res.* **18**, 4123–4130 (1990).
36. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
37. Hudson, T. J. *et al.* A radiation hybrid map of mouse genes. *Nature Genet.* **29**, 201–205 (2001).
38. Van Etten, W. J. *et al.* Radiation hybrid map of the mouse genome. *Nature Genet.* **22**, 384–387 (1999).
39. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22**, 388–393 (1999).
40. Marra, M. *et al.* An encyclopedia of mouse genes. *Nature Genet.* **21**, 191–194 (1999).
41. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
42. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
43. Osoegawa, K. *et al.* Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10**, 116–128 (2000).
44. Gregory, S. G. *et al.* A physical map of the mouse genome. *Nature* **418**, 743–750 (2002).
45. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
46. Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet.* **2**, 573–583 (2001).
47. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6**, 593–608 (1990).
48. Huson, D. H. *et al.* Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics* **17**, S132–S139 (2001).
49. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
50. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
51. Yu, J. *et al.* A draft sequence of the rice genome. *Science* **296**, 79–92 (2002).
52. Battey, J., Jordan, E., Cox, D. & Dove, W. An action plan for mouse genomics. *Nature Genet.* **21**, 73–75 (1999).
53. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279–286 (2001).
54. Zhao, S. *et al.* Mouse BAC ends quality assessment and sequence analyses. *Genome Res.* **11**, 1736–1745 (2001).
55. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
56. Batzoglu, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177–189 (2002).
57. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* (in the press).
58. Mullikin, J. & Ning, Z. The Phusion Assembler. *Genome Res.* (in the press).
59. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
60. Traut, W., Winking, H. & Adolph, S. An extra segment in chromosome 1 of wild *Mus musculus*: a C-band positive homogeneously staining region. *Cytogenet. Cell Genet.* **38**, 290–297 (1984).
61. Weichenhan, D. *et al.* Source and component genes of a 6–200 Mb gene cluster in the house mouse. *Mamm. Genome* **12**, 590–594 (2001).
62. Purmann, L., Plass, C., Gruneberg, M., Winking, H. & Traut, W. A long-range repeat cluster in chromosome 1 of the house mouse, *Mus musculus*, and its relation to a germline homogeneously staining region. *Genomics* **12**, 80–88 (1992).
63. Wong, A. K. & Rattner, J. B. Sequence organization and cytological localization of the minor satellite of mouse. *Nucleic Acids Res.* **16**, 11645–11661 (1988).
64. Joseph, A., Mitchell, A. R. & Miller, O. J. The organization of the mouse satellite DNA at centromeres. *Exp. Cell Res.* **183**, 494–500 (1989).
65. Davisson, M. T. & Roderick, T. H. *Genetic Variants and Strains of the Laboratory Mouse* (eds Lyon, M. F. & Searle, A. G.) 416–427 (Oxford Univ. Press, Oxford, 1989).
66. Mouse Genome Sequencing Consortium Progress in sequencing the mouse genome. *Genesis* **31**, 137–141 (2001).
67. Clark, F. H. Inheritance and linkage relations of mutant characteristics in the deer mouse. *Contrib. Lab. Vert. Biol.* **7**, 1–11 (1938).
68. Castle, W. W. Observations of the occurrence of linkage in rats and mice. *Car. Inst. Wash. Pub.* **288**, 29–36 (1919).
69. Lalley, P. A., Minna, J. D. & Francke, U. Conservation of autosomal gene synteny groups in mouse and man. *Nature* **274**, 160–163 (1978).
70. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
71. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
72. Ohno, S. *Sex Chromosomes and Sex-Linked Genes* (Springer, Berlin, 1996).
73. Sturtevant, A. H. & Beadle, G. W. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* **21**, 554–604 (1936).
74. Ranz, J. M., Casals, F. & Ruiz, A. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11**, 230–239 (2001).
75. Nadeau, J. H. & Sankoff, D. The lengths of undiscovered conserved segments in comparative maps. *Mamm. Genome* **9**, 491–495 (1998).
76. Ferretti, V., Nadeau, J. H. & Sankoff, D. *Combinatorial Pattern Matching, 7th Annual Symposium* (eds Hirschberg, D. & Myers, G.) 159–167 (Springer, Berlin, 1996).
77. Bourque, G. & Pevzner, P. A. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* **12**, 26–36 (2002).
78. Thierry, J. P., Macaya, G. & Bernardi, G. An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* **108**, 219–235 (1976).
79. Salinas, J., Zerial, M., Filipinski, J. & Bernardi, G. Gene distribution and nucleotide sequence organization in the mouse genome. *Eur. J. Biochem.* **160**, 469–478 (1986).
80. Sabeur, G., Macaya, G., Kadi, F. & Bernardi, G. The isochore patterns of mammalian genomes and their phylogenetic implications. *J. Mol. Evol.* **37**, 93–108 (1993).
81. Zerial, M., Salinas, J., Filipinski, J. & Bernardi, G. Gene distribution and nucleotide sequence organization in the human genome. *Eur. J. Biochem.* **160**, 479–485 (1986).
82. Mouchiroud, D., Fichant, G. & Bernardi, G. Compositional compartmentalization and gene composition in the genome of vertebrates. *J. Mol. Evol.* **26**, 198–204 (1987).
83. Mouchiroud, D., Gautier, C. & Bernardi, G. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J. Mol. Evol.* **27**, 311–320 (1988).
84. Mouchiroud, D. & Gautier, C. Codon usage changes and sequence dissimilarity between human and rat. *J. Mol. Evol.* **31**, 81–91 (1990).
85. Robinson, M., Gautier, C. & Mouchiroud, D. Evolution of isochores in rodents. *Mol. Biol. Evol.* **14**, 823–828 (1997).
86. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
87. Mouchiroud, D. *et al.* The distribution of genes in the human genome. *Gene* **100**, 181–187 (1991).
88. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**, 95–102 (1996).
89. Saccone, S., Pavlicek, A., Federico, C., Paces, J. & Bernard, G. Genes, isochores and bands in human chromosomes 21 and 22. *Chromosome Res.* **9**, 533–539 (2001).
90. Bernardi, G. Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1–11 (1986).
91. Bernardi, G., Mouchiroud, D. & Gautier, C. Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.* **28**, 7–18 (1988).
92. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
93. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 2653–2657 (1988).
94. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592 (1962).
95. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
96. Larsen, E., Gunderson, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13**, 1095–1107 (1992).
97. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
98. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
99. Adams, R. L. & Eason, R. Increased G+C content of DNA stabilizes methyl CpG dinucleotides. *Nucleic Acids Res.* **12**, 5869–5877 (1984).
100. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
101. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224**, 149–154 (1969).
102. Kohne, D. E. Evolution of higher-organism DNA. *Q. Rev. Biophys.* **3**, 327–375 (1970).
103. Goodman, M., Barnabas, J., Matsuda, G. & Moore, G. W. Molecular evolution in the descent of man. *Nature* **233**, 604–613 (1971).
104. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808 (2002).
105. Easteal, S., Collet, C. & Betty, D. *The Mammalian Molecular Clock* (Landes, Austin, Texas, 1995).
106. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5**, 182–187 (1996).
107. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90**, 4087–4091 (1993).
108. Bromham, L. Molecular clocks in reptiles: life history influences rate of molecular evolution. *Mol. Biol. Evol.* **19**, 302–309 (2002).
109. Wu, C. I. & Li, W. H. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl Acad. Sci. USA* **82**, 1741–1745 (1985).
110. Smit, A. F., Toth, G., Riggs, A. D. & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417 (1995).
111. Adey, N. B. *et al.* Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol. Biol. Evol.* **11**, 778–789 (1994).

112. Mears, M. L. & Hutchison, C. A. III The evolution of modern lineages of mouse L1 elements. *J. Mol. Evol.* **52**, 51–62 (2001).
113. Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, H. H. Jr A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.* **11**, 1677–1685 (2001).
114. Hardies, S. C. *et al.* LINE-1 (L1) lineages in the mouse. *Mol. Biol. Evol.* **17**, 616–628 (2000).
115. Ohshima, K., Hamada, M., Terai, Y. & Okada, N. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol. Cell Biol.* **16**, 3756–3764 (1996).
116. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748 (1996).
117. Quentin, Y. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Res.* **22**, 2222–2227 (1994).
118. Kim, J. & Deininger, P. L. Recent amplification of rat ID sequences. *J. Mol. Biol.* **261**, 322–327 (1996).
119. Lee, I. Y. *et al.* Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome Res.* **8**, 1022–1037 (1998).
120. Serdobova, I. M. & Kramerov, D. A. Short retroposons of the B2 superfamily: evolution and application for the study of rodent phylogeny. *J. Mol. Evol.* **46**, 202–214 (1998).
121. Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds) *Retroviruses* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
122. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR- transposons. *Nucleic Acids Res.* **21**, 1863–1872 (1993).
123. Hamilton, B. A. & Frankel, W. N. Of mice and genome sequence. *Cell* **107**, 13–16 (2001).
124. Turner, G. *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531–1535 (2001).
125. Kidwell, M. G. Horizontal transfer. *Curr. Opin. Genet. Dev.* **2**, 868–873 (1992).
126. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
127. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA* **94**, 1872–1877 (1997).
128. Bernardi, G. The isochore organization of the human genome. *Annu. Rev. Genet.* **23**, 637–661 (1989).
129. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51**, 17–37 (1992).
130. Korenberg, J. R. & Rykowski, M. C. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400 (1988).
131. Boyle, A. L., Ballard, S. G. & Ward, D. C. Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence *in situ* hybridization. *Proc. Natl Acad. Sci. USA* **87**, 7757–7761 (1990).
132. Lyon, M. F. X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
133. Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl Acad. Sci. USA* **97**, 6634–6639 (2000).
134. Boissinot, S. & Furano, A. V. Adaptive evolution in LINE-1 retrotransposons. *Mol. Biol. Evol.* **18**, 2186–2194 (2001).
135. Beckman, J. S. & Weber, J. L. Survey of human and rat microsatellites. *Genomics* **12**, 627–631 (1992).
136. Toth, G., Gaspari, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981 (2000).
137. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95**, 10774–10778 (1998).
138. Santibanez-Koref, M. F., Gangeswaran, R. & Hancock, J. M. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol. Biol. Evol.* **18**, 2119–2123 (2001).
139. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
140. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
141. Roest Croliius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
142. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
143. Kulp, D., Haussler, D., Reese, M. G. & Eckman, F. H. Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.* 232–244 (1997).
144. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
145. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
146. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
147. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
148. Daly, M. J. Estimating the human gene count. *Cell* **109**, 283–284 (2002).
149. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
150. The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
151. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
152. Steinle, V. *et al.* A novel DNA-binding regulatory factor is mutated in primary MHC class II deficiency (bare lymphocyte syndrome). *Genes Dev.* **9**, 1021–1032 (1995).
153. Sun, H., Tsunenari, T., Yau, K. W. & Nathans, J. The vitelliform macular dystrophy protein defines a new family of chloride channels. *Proc. Natl Acad. Sci. USA* **99**, 4008–4013 (2002).
154. Yasunaga, S. *et al.* A mutation in OTOF, encoding otoferlin, a FER-1-like protein, causes DFNB9, a nonsyndromic form of deafness. *Nature Genet.* **21**, 363–369 (1999).
155. den Hollander, A. I. *et al.* Leber congenital amaurosis and retinitis pigmentosa with Coats-like exudative vasculopathy are associated with mutations in the crumbs homologue 1 (CRB1) gene. *Am. J. Hum. Genet.* **69**, 198–203 (2001).
156. den Hollander, A. I. *et al.* Mutations in a human homologue of *Drosophila* crumbs cause retinitis pigmentosa (RP12). *Nature Genet.* **23**, 217–221 (1999).
157. Maeda, N. *et al.* Diet-induced insulin resistance in mice lacking adiponectin/ACRP30. *Nature Med.* **8**, 731–737 (2002).
158. Clausen, B. E. *et al.* Residual MHC class II expression on mature dendritic cells and activated B cells in RFX5-deficient mice. *Immunity* **8**, 143–155 (1998).
159. Garcia-Meunier, P., Etienne-Julan, M., Fort, P., Piechaczky, M. & Bonhomme, F. Concerted evolution in the GAPDH family of retrotransposed pseudogenes. *Mamm. Genome* **4**, 695–703 (1993).
160. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, S140–S148 (2001).
161. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. & Guigo, R. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11**, 1574–1583 (2001).
162. Alexandersson, M., Cawley, S. & Pachter, L. SLAM—cross-species GeneFinding and alignment with a generalized pair hidden Markov model. *Genome Res.* (in the press).
163. Raymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420**, 582–586 (2002).
164. Blake, D. J., Weir, A., Newey, S. E. & Davies, K. E. Function and genetics of dystrophin-related proteins in muscle. *Physiol. Rev.* **82**, 291–329 (2002).
165. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929 (2001).
166. Storz, G. An expanding universe of noncoding RNAs. *Science* **296**, 1260–1263 (2002).
167. Eddy, S. R. Computational genomics of noncoding RNA genes. *Cell* **109**, 137–140 (2002).
168. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
169. Daniels, G. R. & Deininger, P. L. Repeat sequence families derived from mammalian tRNA genes. *Nature* **317**, 819–822 (1985).
170. Lawrence, C., McDonnell, D. & Ramsey, W. Analysis of repetitive sequence elements containing tRNA-like sequences. *Nucleic Acids Res.* **13**, 4239–4252 (1985).
171. Baron, C. & Bock, A. *tRNA: Structure, Biosynthesis, and Function* (eds Soll, D. & RajBhandary, U. L.) 529–544 (Am. Soc. Microbiol., Washington DC, 1995).
172. Crick, F. H. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
173. Guthrie, C. & Abelson, J. *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds Strathern, J. N., Jones, E. W. & Broach, J. R.) 487–528 (Cold Spring Harbor Laboratory Press, Woodbury, New York, 1982).
174. Ponting, C. P. & Russell, R. R. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 45–71 (2002).
175. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**, 1048–1059 (2002).
176. Ponting, C. P., Mott, R., Bork, P. & Copley, R. R. Novel protein domains and repeats in *Drosophila melanogaster*: insights into structure, function, and evolution. *Genome Res.* **11**, 1996–2008 (2001).
177. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
178. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
179. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
180. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
181. Makalowski, W. & Boguski, M. S. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J. Mol. Evol.* **47**, 119–121 (1998).
182. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
183. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
184. Nekrutenko, A., Makova, K. D. & Li, W. H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12**, 198–202 (2002).
185. Sharp, P. M. In search of molecular darwinism. *Nature* **385**, 111–112 (1997).
186. Letunic, I. *et al.* Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30**, 242–244 (2002).
187. Mott, R., Schultz, J., Bork, P. & Ponting, C. P. Predicting protein cellular localization using a domain projection method. *Genome Res.* **12**, 1168–1174 (2002).
188. Hurst, L. D. & Smith, N. G. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
189. Goodstadt, L. & Ponting, C. P. Sequence variation and disease in the wake of the draft human genome. *Hum. Mol. Genet.* **10**, 2209–2214 (2001).
190. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
191. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
192. Fredman, D. *et al.* HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.* **30**, 387–391 (2002).
193. Young, J. M. *et al.* Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* **11**, 535–546 (2002).
194. Zhang, X. & Firestein, S. The olfactory receptor gene superfamily of the mouse. *Nature Neurosci.* **5**, 124–133 (2002).
195. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. The complete human olfactory subgenome. *Genome Res.* **11**, 685–702 (2001).
196. Rouquier, S. *et al.* Distribution of olfactory receptor genes in the human genome. *Nature Genet.* **18**, 243–250 (1998).
197. Del Punta, K. *et al.* Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. *Nature* **419**, 70–74 (2002).
198. Nelson, D. R. Cytochrome P450 and the individuality of species. *Arch. Biochem. Biophys.* **369**, 1–10 (1999).

199. Lane, R. P. *et al.* Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl Acad. Sci. USA* **98**, 7390–7395 (2001).
200. Rossant, J. & Cross, J. C. Placental development: lessons from mouse mutants. *Nature Rev. Genet.* **2**, 538–548 (2001).
201. Georgiades, P., Ferguson-Smith, A. C. & Burton, G. J. Comparative developmental anatomy of the murine and human definitive placenta. *Placenta* **23**, 3–19 (2002).
202. Deussing, J. *et al.* Identification and characterization of a dense cluster of placenta-specific cysteine peptidase genes and related genes on mouse chromosome 13. *Genomics* **79**, 225–240 (2002).
203. Afonso, S., Tovar, C., Romagnano, L. & Babiarz, B. Control and expression of cystatin C by mouse decidua cultures. *Mol. Reprod. Dev.* **61**, 155–163 (2002).
204. Sutton, K. A. & Wilkinson, M. F. The rapidly evolving Pem homeobox gene and Agtr2, Ant2, and Lamp2 are closely linked in the proximal region of the mouse X chromosome. *Genomics* **45**, 447–450 (1997).
205. Wilkinson, M. F., Kleeman, J., Richards, J. & MacLeod, C. L. A novel oncofetal gene is expressed in a stage-specific manner in murine embryonic development. *Dev. Biol.* **141**, 451–455 (1990).
206. Han, Y. J., Park, A. R., Sung, D. Y. & Chun, J. Y. Pxx, a novel murine homeobox gene expressed in placenta. *Gene* **207**, 159–166 (1998).
207. Chun, J. Y., Han, Y. J. & Ahn, K. Y. Pxx homeobox gene is X-linked and specifically expressed in trophoblast cells of mouse placenta. *Dev. Dyn.* **216**, 257–266 (1999).
208. Takasaki, N., McIsaac, R. & Dean, J. Gpbox (Pxx2), a homeobox gene preferentially expressed in female germ cells at the onset of sexual dimorphism in mice. *Dev. Biol.* **223**, 181–193 (2000).
209. Lundwall, A. & Lazure, C. A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon. *FEBS Lett.* **374**, 53–56 (1995).
210. Simon, A. M., Veyssiere, G. & Jean, C. Structure and sequence of a mouse gene encoding an androgen-regulated protein: a new member of the seminal vesicle secretory protein family. *J. Mol. Endocrinol.* **15**, 305–316 (1995).
211. Morel, L. *et al.* Mouse seminal vesicle secretory protein of 99 amino acids (MSVSP99): characterization and hormonal and developmental regulation. *J. Androl.* **22**, 549–557 (2001).
212. Linzer, D. I. & Fisher, S. J. The placenta and the prolactin family of hormones: regulation of the physiology of pregnancy. *Mol. Endocrinol.* **13**, 837–840 (1999).
213. Huang, Y. H., Chu, S. T. & Chen, Y. H. A seminal vesicle autoantigen of mouse is able to suppress sperm capacitation-related events stimulated by serum albumin. *Biol. Reprod.* **63**, 1562–1566 (2000).
214. Yoshida, M., Kaneko, M., Kurachi, H. & Osawa, M. Identification of two rodent genes encoding homologues to seminal vesicle autoantigen: a gene family including the gene for prolactin-inducible protein. *Biochem. Biophys. Res. Commun.* **281**, 94–100 (2001).
215. Bain, P. A., Yoo, M., Clarke, T., Hammond, S. H. & Payne, A. H. Multiple forms of mouse 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4 isomerase and differential expression in gonads, adrenal glands, liver, and kidneys of both sexes. *Proc. Natl Acad. Sci. USA* **88**, 8870–8874 (1991).
216. Payne, A. H., Abbaszade, I. G., Clarke, T. R., Bain, P. A. & Park, C. H. The multiple murine 3 beta-hydroxysteroid dehydrogenase isoforms: structure, function, and tissue- and developmentally specific expression. *Steroids* **62**, 169–175 (1997).
217. Blume, N. *et al.* Characterization of Cyp2d22, a novel cytochrome P450 expressed in mouse mammary cells. *Arch. Biochem. Biophys.* **381**, 191–204 (2000).
218. Lakso, M., Masaki, R., Noshiro, M. & Negishi, M. Structures and characterization of sex-specific mouse cytochrome P-450 genes as members within a large family. Duplication boundary and evolution. *Eur. J. Biochem.* **195**, 477–486 (1991).
219. Tegoni, M. *et al.* Mammalian odorant binding proteins. *Biochim. Biophys. Acta* **1482**, 229–240 (2000).
220. Miyawaki, A., Matsushita, F., Ryo, Y. & Mikoshiba, K. Possible pheromone-carrier function of two lipocalin proteins in the vomeronasal organ. *EMBO J.* **13**, 5835–5842 (1994).
221. Karn, R. C. & Nachman, M. W. Reduced nucleotide variability at an androgen-binding protein locus (Abpa) in house mice: evidence for positive natural selection. *Mol. Biol. Evol.* **16**, 1192–1197 (1999).
222. Karn, R. C., Orth, A., Bonhomme, F. & Boursot, P. The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. *Mol. Biol. Evol.* **19**, 462–471 (2002).
223. Singer, A. G., Macrides, F., Clancy, A. N. & Agosta, W. C. Purification and analysis of a proteinaceous aphrodisiac pheromone from hamster vaginal discharge. *J. Biol. Chem.* **261**, 13323–13326 (1986).
224. Zhang, J., Dyer, K. D. & Rosenberg, H. F. Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl Acad. Sci. USA* **97**, 4701–4706 (2000).
225. Natarajan, K., Dimasi, N., Wang, J., Margulies, D. H. & Mariuzza, R. A. MHC class I recognition by Ly49 natural killer cell receptors. *Mol. Immunol.* **38**, 1023–1027 (2002).
226. Natarajan, K., Dimasi, N., Wang, J., Mariuzza, R. A. & Margulies, D. H. Structure and function of natural killer cell receptors: multiple molecular solutions to self, nonself discrimination. *Annu. Rev. Immunol.* **20**, 853–885 (2002).
227. Yeager, M. & Hughes, A. L. Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. *Immunol. Rev.* **167**, 45–58 (1999).
228. Ichikawa, T., Itakura, T. & Negishi, M. Functional characterization of two cytochrome P-450s within the mouse, male-specific steroid 16 alpha-hydroxylase gene family: expression in mammalian cells and chimeric proteins. *Biochemistry* **28**, 4779–4784 (1989).
229. Miao, Y. J., Subramaniam, N. & Carlson, D. M. cDNA cloning and characterization of rat salivary glycoproteins. Novel members of the proline-rich-protein multigene families. *Eur. J. Biochem.* **228**, 343–350 (1995).
230. Whelan, S., Lio, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**, 262–272 (2001).
231. Taveré, S. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lec. Math. Life Sci.* **17**, 57–86 (1986).
232. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**, 815–824 (1999).
233. Suzuki, Y. *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).
234. Kozak, M. Do the 5' untranslated domains of human cDNAs challenge the rules for initiation of translation (or is it vice versa)? *Genomics* **70**, 396–406 (2000).
235. Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* **63**, 405–445 (1999).
236. Batzoglu, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
237. Ogata, H., Fujibuchi, W. & Kanehisa, M. The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* **390**, 99–103 (1996).
238. Burge, C. B., Padgett, R. A. & Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2**, 773–785 (1998).
239. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
240. Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**, 832–839 (2002).
241. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566 (2001).
242. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
243. Dermitzakis, E. & Clark, A. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
244. Ooi, G. T., Hurst, K. R., Poy, M. N., Rechler, M. M. & Boisclair, Y. R. Binding of STAT5a and STAT5b to a single element resembling a gamma-interferon-activated sequence mediates the growth hormone induction of the mouse acid-labile subunit promoter in liver cells. *Mol. Endocrinol.* **12**, 675–687 (1998).
245. Suwanichkul, A., Boisclair, Y. R., Olne, R. C., Durham, S. K. & Powell, D. R. Conservation of a growth hormone-responsive promoter element in the human and mouse acid-labile subunit genes. *Endocrinology* **141**, 833–838 (2000).
246. Campbell, S. M., Rosen, J. M., Hennighausen, L. G., Strech-Jurk, U. & Sippel, A. E. Comparison of the whey acidic protein genes of the rat and mouse. *Nucleic Acids Res.* **12**, 8685–8697 (1984).
247. Dermitzakis, E. T. *et al.* Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582 (2002).
248. Koop, B. F. Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.* **11**, 367–371 (1995).
249. DeBry, R. W. & Seldin, M. F. Human/mouse homology relationships. *Genomics* **33**, 337–351 (1996).
250. Gottgens, B. *et al.* Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res.* **11**, 87–97 (2001).
251. Shiraishi, T. *et al.* Sequence conservation at human and mouse orthologous common fragile regions, FRA3B/FHIT and Fra14A2/Fhit. *Proc. Natl Acad. Sci. USA* **98**, 5722–5727 (2001).
252. Wilson, M. D. *et al.* Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.* **29**, 1352–1365 (2001).
253. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
254. Chiaromonte, F. *et al.* Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl Acad. Sci. USA* **98**, 14503–14508 (2001).
255. Matassi, G., Sharp, P. M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791 (1999).
256. Williams, E. J. & Hurst, L. D. The proteins of linked genes evolve at similar rates. *Nature* **407**, 900–903 (2000).
257. Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**, 481–489 (2001).
258. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340 (2002).
259. Castresana, J. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res.* **30**, 1751–1756 (2002).
260. Smith, N. G. C., Webster, M. & Ellegren, H. Deterministic mutation rate variation in the human genome. *Genome Res.* **12**, 1350–1356 (2002).
261. Hardison, R. *et al.* Co-variation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome Res.* (in the press).
262. Bernardi, G. The human genome: organization and evolutionary history. *Ann. Rev. Genet.* **23**, 637–661 (1995).
263. Hurst, L. D. & Williams, E. J. B. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene* **261**, 107–114 (2000).
264. Bernardi, G. Misunderstandings about isochores. Part 1. *Gene* **276**, 3–13 (2001).
265. The SNP Consortium An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
266. Perry, J. & Ashworth, A. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9**, 987–989 (1999).
267. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
268. Nachman, M. W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001).
269. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
270. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**, 213–227 (1994).
271. Hudson, R. R. & Kaplan, N. L. Deleterious background selection with recombination. *Genetics* **141**, 1605–1617 (1995).
272. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
273. Birky, C. W. & Walsh, J. B. Effects of linkage on rates of molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 6414–6418 (1988).
274. Francino, M. P. & Ochman, H. Strand asymmetries in DNA evolution. *Trends Genet.* **13**, 240–245 (1997).

275. Gilbert, N., Lutz-Prigge, S. & Moran, J. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315–325 (2002).
276. Symer, D. *et al.* Human I1 retrotransposition is associated with genetic instability *in vivo*. *Cell* **110**, 327–338 (2002).
277. Moran, J. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
278. Hughes, J. F. & Coffin, J. M. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature Genet.* **29**, 487–489 (2001).
279. Wolfe, K. H. Mammalian DNA replication: mutation biases and the mutation rate. *J. Theor. Biol.* **149**, 441–451 (1991).
280. Gu, X. & Li, W. H. A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. *J. Mol. Evol.* **38**, 468–475 (1994).
281. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
282. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
283. Wade, C. M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574–578 (2002).
284. Loftus, S. K. *et al.* Mutation of melanosome protein RAB38 in chocolate mice. *Proc. Natl Acad. Sci. USA* **99**, 4471–4476 (2002).
285. Paigen, K. & Eppig, J. T. A mouse phenome project. *Mamm. Genome* **11**, 715–717 (2000).
286. Doerge, R. W. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Rev. Genet.* **3**, 43–52 (2002).
287. Cormier, R. T. *et al.* The Mom1AKR intestinal tumour resistance region consists of Pla2g2a and a locus distal to D4Mit64. *Oncogene* **19**, 3182–3192 (2000).
288. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* **10**, 1126–1137 (2000).
289. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnol.* **18**, 1001–1005 (2000).
290. Heiskanen, M. *et al.* CGH, cDNA and tissue microarray analyses implicate FGFR2 amplification in a small subset of breast tumors. *Anal. Cell Pathol.* **22**, 229–234 (2001).
291. Cai, W. W. *et al.* Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nature Biotechnol.* **20**, 393–396 (2002).
292. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
293. Mikkers, H. *et al.* High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nature Genet.* **32**, 153–159 (2002).
294. Hwang, H. C. *et al.* Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis. *Proc. Natl Acad. Sci. USA* **99**, 11293–11298 (2002).
295. Lund, A. *et al.* Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nature Genet.* **32**, 160–165 (2002).
296. Janne, P. A. *et al.* Functional overlap between murine Inpp5b and Ocr1l may explain why deficiency of the murine ortholog for Ocr1l does not cause Lowe syndrome in mice. *J. Clin. Invest.* **101**, 2042–2053 (1998).
297. Saitou, N. & Nei, M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
298. Sokal, R. & Rohlf, F. *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York, 1995).
299. Sutton, K. A. & Wilkinson, M. F. Rapid evolution of a homeodomain: evidence for positive selection. *J. Mol. Evol.* **45**, 579–588 (1997).
300. Kasper, S. & Matusik, R. J. Rat probasin: structure and function of an outlier lipocalin. *Biochim. Biophys. Acta* **1482**, 249–258 (2000).
301. Briand, L. *et al.* Odorant and pheromone binding by aphrodisin, a hamster aphrodisiac protein. *FEBS Lett.* **476**, 179–185 (2000).
302. Gow, A. *et al.* CNS myelin and sertoli cell tight junction strands are absent in Osp/claudin-11 null mice. *Cell* **99**, 649–659 (1999).
303. Kollmar, R., Nakamura, S. K., Kappler, J. A. & Hudspeth, A. J. Expression and phylogeny of claudins in vertebrate primordia. *Proc. Natl Acad. Sci. USA* **98**, 10196–10201 (2001).
304. Ashcroft, G. S. *et al.* Secretory leukocyte protease inhibitor mediates non-redundant functions necessary for normal wound healing. *Nature Med.* **6**, 1147–1153 (2000).
305. Henderson, C. J., Bammler, T. & Wolf, C. R. Deduced amino acid sequence of a murine cytochrome P-450 Cyp4a protein: developmental and hormonal regulation in liver and kidney. *Biochim. Biophys. Acta* **1200**, 182–190 (1994).
306. Simpson, A. E. The cytochrome P450 4 (CYP4) family. *Gen. Pharmacol.* **28**, 351–359 (1997).
307. Sundseth, S. S. & Waxman, D. J. Sex-dependent expression and clofibrate inducibility of cytochrome P450 4A fatty acid omega-hydroxylases. Male specificity of liver and kidney CYP4A2 mRNA and tissue-specific regulation by growth hormone and testosterone. *J. Biol. Chem.* **267**, 3915–3921 (1992).
308. Myal, Y. *et al.* Tissue-specific androgen-inhibited gene expression of a submaxillary gland protein, a rodent homolog of the human prolactin-inducible protein/GCDFP-15 gene. *Endocrinology* **135**, 1605–1610 (1994).
309. Huang, Y. H., Chu, S. T. & Chen, Y. H. Seminal vesicle autoantigen, a novel phospholipid-binding protein secreted from luminal epithelium of mouse seminal vesicle, exhibits the ability to suppress mouse sperm motility. *Biochem. J.* **343**, 241–248 (1999).
310. Ann, D. K., Smith, M. K. & Carlson, D. M. Molecular evolution of the mouse proline-rich protein multigene family. Insertion of a long interspersed repeated DNA element. *J. Biol. Chem.* **263**, 10887–10893 (1988).
311. Rosinski-Chupin, I. & Rougeon, F. A new member of the glutamine-rich protein gene family is characterized by the absence of internal repeats and the androgen control of its expression in the submandibular gland of rats. *J. Biol. Chem.* **265**, 10709–10713 (1990).
312. Rajkovic, A., Yan, C., Yan, W., Klysk, M. & Matzuk, M. M. Obox, a family of homeobox genes preferentially expressed in germ cells. *Genomics* **79**, 711–717 (2002).
313. Talley, H. M., Laukaitis, C. M. & Karn, R. C. Female preference for male saliva: implications for sexual isolation of *Mus musculus* subspecies. *Evol. Int. J. Org. Evol.* **55**, 631–634 (2001).
314. Dlouhy, S. R., Taylor, B. A. & Karn, R. C. The genes for mouse salivary androgen-binding protein (ABP) subunits alpha and gamma are located on chromosome 7. *Genetics* **115**, 535–543 (1987).
315. Jia, H. P. *et al.* A novel murine beta-defensin expressed in tongue, esophagus, and trachea. *J. Biol. Chem.* **275**, 33314–33320 (2000).
316. Peters, J. Nonspecific esterases of *Mus musculus*. *Biochem. Genet.* **20**, 585–606 (1982).
317. Abou-Haila, A., Orgebin-Crist, M. C., Skudlarek, M. D. & Tulsiani, D. R. Identification and androgen regulation of egasyn in the mouse epididymis. *Biochim. Biophys. Acta* **1401**, 177–186 (1998).
318. Lin, J., Toft, D. J., Bengtson, N. W. & Linzer, D. I. Placental prolactins and the physiology of pregnancy. *Recent Prog. Horm. Res.* **55**, 37–51 (2000).
319. Goffin, V., Binart, N., Touraine, P. & Kelly, P. A. Prolactin: the new biology of an old hormone. *Annu. Rev. Physiol.* **64**, 47–67 (2002).
320. Batten, D., Dyer, K. D., Domachowski, J. B. & Rosenberg, H. F. Molecular cloning of four novel murine ribonuclease genes: unusual expansion within the ribonuclease A gene family. *Nucleic Acids Res.* **25**, 4235–4239 (1997).
321. Cormier, S. A. *et al.* Mouse eosinophil-associated ribonucleases: a unique subfamily expressed during hematopoiesis. *Mamm. Genome* **12**, 352–361 (2001).
322. Tsui, F. W. *et al.* Molecular characterization and mapping of murine genes encoding three members of the stefin family of cysteine proteinase inhibitors. *Genomics* **15**, 507–514 (1993).
323. Parham, P. Virtual reality in the MHC. *Immunol. Rev.* **167**, 5–15 (1999).
324. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
325. Flicek, P. *et al.* Leveraging the mouse genome for gene prediction in human: From the whole-genome shotgun reads to a global synteny map. *Genome Res.* (in the press).
326. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* (in the press).
327. Guigó, R. *et al.* Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl Acad. Sci. USA* (in the press).
328. Schwartz, S. *et al.* Human-mouse alignments with Blastz. *Genome Res.* (in the press).
329. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* (in the press).
330. Roskin, K. M. Score Functions for Assessing Conservation in Locally Aligned Regions of DNA from Two Species. UCSC Tech Report UCSC-CRL-02-30, School of Engineering, Univ. California (2002).

Supplementary Information accompanies the paper on *Nature's* website
(<http://www.nature.com/nature>).

Acknowledgements We thank J. Takahashi and M. Johnston for comments on the manuscript; the Mouse Liaison Group for strategic advice; L. Gaffney, D. Leja and K.-S. Toh for graphical help; B. Graham and G. Roberts for administrative work on sequencing of individual mouse BACs; and P. Kassos and M. McMurtry for secretarial assistance. We thank D. Hill and L. Corbani of the Mouse Genome Informatics Group for their contributions to the GO analysis for mouse and human, and the members of the Bork group at EMBL for discussions. Funding was provided by the National Institutes of Health (National Human Genome Research Institute, National Cancer Institute, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of General Medical Sciences, National Eye Institute, National Institute of Environmental Health Sciences, National Institute of Aging, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institute on Deafness and Other Communication Disorders, National Institute of Mental Health, National Institute on Drug Abuse, National Center for Research Resources, the National Heart Lung and Blood Institute and The Fogarty International Center); the Wellcome Trust; the Howard Hughes Medical Institute; the United States Department of Energy; the National Science Foundation; the Medical Research Council; NSERC; BMBF (German Ministry for Research and Education); the European Molecular Biology Laboratory; Plan Nacional de I+D and Instituto Carlos III; Swiss National Science Foundation, NCCR Frontiers in Genetics, the Swiss Cancer League and the 'Childcare' and 'J. Lejeune' Foundations; and the Ministry of Education, Culture, Sports, Science and Technology of Japan. The initial threefold sequence coverage was partly supported by the Mouse Sequencing Consortium (GlaxoSmithKline, Merck and Affymetrix) through the Foundation for the National Institutes of Health. We acknowledge A. Holden for coordinating the Mouse Sequencing Consortium. We thank the Sanger Institute systems group for maintenance and provision of the computer resource. The MGSC also used Hewlett-Packard Company's BioCluster, a configuration of 27 HP AlphaServer ES40 systems with 100 CPUs and 1 terabyte of storage. The BioCluster is housed in Hewlett-Packard's IQ Solutions Center, and was accessed remotely. The computing resource greatly accelerated the analysis.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H.W. (e-mail: waterston@gs.washington.edu), K.L.T. (e-mail: kersli@genome.wi.mit.edu) or E.S.L. (e-mail: lander@genome.wi.mit.edu).

Authors' contributions The following authors contributed to project leadership: R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, M. R. Brent, F. S. Collins, R. Guigó, R. C. Hardison, D. Haussler, D. B. Jaffe, W. J. Kent, W. Miller, C. P. Ponting, A. Smit, M. C. Zody and E. S. Lander.

Robert H. Waterston^{1*}, Kerstin Lindblad-Toh^{2*}, Ewan Birney^{3*}, Jane Rogers⁴, Josep F. Abril^{5*}, Pankaj Agarwal^{6*}, Richa Agarwala⁷, Rachel Ainscough⁴, Marina Alexandersson^{8*}, Peter An², Stylianos E. Antonarakis^{9*}, John Attwood⁴, Robert Baertsch^{10*}, Jonathon Bailey⁴, Karen Barlow⁴, Stephan Beck⁴, Eric Berry^{2*}, Bruce Birren², Toby Bloom², Peer Bork^{11*}, Marc Botcherby¹², Nicolas Bray^{13*}, Michael R. Brent^{14*}, Daniel G. Brown^{2,15*}, Stephen D. Brown¹², Carol Bult^{16*}, John Burton⁴, Jonathan Butler^{2*}, Robert D. Campbell¹², Piero Carninci¹⁷, Simon Cawley^{18*}, Francesca Chiaromonte^{19*}, Asif T. Chinwalla^{1*}, Deanna M. Church^{7*}, Michele Clamp^{4*}, Christopher Clee⁴, Francis S. Collins^{20*}, Lisa L. Cook¹, Richard R. Copley^{21*}, Alan Coulson⁴, Olivier Couronne^{13*}, James Cuff^{4*}, Val Curwen^{4*}, Tim Cutts^{4*}, Mark Daly^{2*}, Robert David², Joy Davies⁴, Kimberly D. Delehaunty¹, Justin Deri², Emmanouil T. Dermitzakis^{9*}, Colin Dewey^{22*}, Nicholas J. Dickens^{23*}, Mark Diekhans^{10*}, Sheila Dodge², Inna Dubchak^{13*}, Diane M. Dunn²⁴, Sean R. Eddy^{25*}, Laura Elnitski^{26*}, Richard D. Emes^{23*}, Pallavi Eswara^{27*}, Eduardo Eyras^{4*}, Adam Felsenfeld^{20*}, Ginger A. Fewell¹, Paul Flicek^{14*}, Karen Foley², Wayne N. Frankel^{16*}, Lucinda A. Fulton^{1*}, Robert S. Fulton¹, Terrence S. Furey^{10*}, Diane Gage², Richard A. Gibbs²⁸, Gustavo Glusman^{29*}, Sante Gnerre^{2*}, Nick Goldman^{3*}, Leo Goodstadt^{23*}, Darren Grafham⁴, Tina A. Graves¹, Eric D. Green^{30*}, Simon Gregory^{4*}, Roderic Guigo^{5*}, Mark Guyer²⁰, Ross C. Hardison^{31*}, David Haussler^{32*}, Yoshihide Hayashizaki¹⁷, LaDeana W. Hillier^{1*}, Angela Hinrichs^{10*}, Wratko Hlavina^{7*}, Timothy Holzer², Fan Hsu^{10*}, Axin Hua³³, Tim Hubbard^{4*}, Adrienne Hunt⁴, Ian Jackson¹², David B. Jaffe^{2*}, L. Steven Johnson²⁵, Matthew Jones⁴, Thomas A. Jones²⁵, Ann Joy⁴, Michael Kamal^{2*}, Elinor K. Karlsson^{2*}, Donna Karolchik^{10*}, Arkadiusz Kasprzyk^{3*}, Jun Kawai¹⁷, Evan Keibler^{14*}, Cristyn Kells², W. James Kent^{10*}, Andrew Kirby^{2*}, Diana L. Kolbe^{26*}, Ian Korf^{14*}, Raju S. Kucherlapati³⁴, Edward J. Kulbokas III^{2*}, David Kulp^{18*}, Tom Landers², J. P. Leger², Steven Leonard⁴, Ivica Letunic^{11*}, Rosie Levine², Jia Li^{35*}, Ming Li^{36*}, Christine Lloyd⁴, Susan Lucas³⁷, Bin Ma^{38*}, Donna R. Maglott^{7*}, Elaine R. Mardis¹, Lucy Matthews⁴, Evan Mauceli^{2*}, John H. Mayer², Megan McCarthy², W. Richard McCombie³⁹, Stuart McLaren⁴, Kirsten McLay⁴, John D. McPherson¹, Jim Meldrim², Beverley Meredith⁴, Jill P. Mesirov^{2*}, Webb Miller^{27*}, Tracie L. Miner¹, Emmanuel Mongin³, Kate T. Montgomery³⁴, Michael Morgan⁴⁰, Richard Mott^{21*}, James C. Mullikin^{4*}, Donna M. Muzny²⁸, William E. Nash¹, Joanne O. Nelson¹, Michael N. Nhan¹, Robert Nicol², Zemin Ning^{4*}, Chad Nusbaum², Michael J. O'Connor²⁷, Yasushi Okazaki¹⁷, Karen Oliver⁴, Emma Overton-Larty⁴, Lior Pachter^{6*}, Genis Parra^{5*}, Kymberlie H. Pepin¹, Jane Peterson²⁰, Pavel Pevzner^{41*}, Robert Plumb⁴, Craig S. Pohl¹, Alex Poliakov^{13*}, Tracy C. Ponce¹, Chris P. Ponting^{23*}, Simon Potter^{4*}, Michael Quail⁴, Alexandre Reymond^{9*}, Bruce A. Roe³³, Krishna M. Roskin^{10*}, Edward M. Rubin¹³, Alistair G. Rust³, Ralph Santos², Victor Sapozhnikov^{7*}, Brian Schultz¹, Jörg Schultz^{42*}, Matthias S. Schwartz^{10*}, Scott Schwartz^{27*}, Carol Scott⁴, Steven Seaman², Steve Searle^{4*}, Ted Sharpe², Andrew Sheridan², Ratna Shownkeen⁴, Sarah Sims⁴, Jonathan B. Singer^{2*}, Guy Slater^{3*}, Arian Smit^{29*}, Douglas R. Smith⁴³, Brian Spencer², Arne Stabenau^{3*}, Nicole Stange-Thomann², Charles Sugnet^{10*}, Mikita Suyama^{11*}, Glenn Tesler^{41*}, Johanna Thompson¹, David Torrents^{11*}, Evanne Trevaskis¹, John Tromp^{44*}, Catherine Ucla^{9*}, Abel Ureta-Vidal³, Jade P. Vinson^{2*}, Andrew C. von Niederhausern²⁴, Claire M. Wade^{2*}, Melanie Wall⁴, Ryan J. Weber^{10*}, Robert B. Weiss²⁴, Michael C. Wendl¹, Anthony P. West⁴, Kris Wetterstrand²⁰, Raymond Wheeler^{18*}, Simon Whelan^{3*}, Jamey Wierzbowski², David Willey⁴, Sophie Williams⁴, Richard K. Wilson¹, Eitan Winter^{23*}, Kim C. Worley^{45*}, Dudley Wyman², Shan Yang³¹, Shiaw-Pyng Yang^{1*}, Evgeny M. Zdobnov^{11*}, Michael C. Zody^{2*} & Eric S. Lander^{2,46*}

Affiliations for authors: 1, Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA; 2, Whitehead Institute/MIT Center for Genome Research, 320 Charles Street, Cambridge, Massachusetts 02141, USA; 3, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 4, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; 5, Research Group in Biomedical Informatics, Institut Municipal d'Investigació, Medica/Universitat Pompeu Fabra, Centre de Regulació Genómica, Barcelona, Catalonia, Spain; 6, Bioinformatics, GlaxoSmithKline, UW2230, 709 Swedeland Road, King of Prussia, Pennsylvania 19406, USA; 7, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20892, USA; 8, Department of Mathematics, University of California at Berkeley, 970 Evans Hall, Berkeley, California 94720, USA; 9, Division of Medical Genetics, University of Geneva Medical School, 1 rue Michel-Servet, CH-1211 Geneva, Switzerland; 10, Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; 11, EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany; 12, UK MRC Mouse Sequencing Consortium, MRC Mammalian Genetics Unit, Harwell OX11 0RD, UK; 13, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mailstop 84-171, Berkeley, California 94720, USA; 14, Department of Computer Science, Washington University, Box 1045, St Louis, Missouri 63130, USA; 15, School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada; 16, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA; 17, Laboratory for Genome Exploration, RIKEN Genomic Sciences Center, Yokohama Institute, 1-7-22 Suchiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan; 18, Affymetrix Inc., Emeryville, California 94608, USA; 19, Departments of Statistics and Health Evaluation Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 20, National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Room 4B09, Bethesda, Maryland 20892, USA; 21, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK; 22, Department of Electrical Engineering, University of California, Berkeley, 231 Cory Hall, Berkeley, California 94720, USA; 23, Department of Human Anatomy and Genetics, MRC Functional Genetics Unit, University of Oxford, South Parks Road, Oxford OX1 3QX, UK; 24, Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA; 25, Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63110, USA; 26, Departments of Biochemistry and Molecular Biology and Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 27, Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 28, Baylor College of Medicine, Human Genome Sequencing Center, One Baylor Plaza, MSC-226, Houston, Texas 77030, USA; 29, The Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA; 30, National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50, Room 5523, Bethesda, Maryland 20892, USA; 31, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 32, Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA; 33, Department of Chemistry and Biochemistry, University of Oklahoma Advanced Center for Genome Technology, University of Oklahoma, 620 Parrington Oval, Room 311, Norman, Oklahoma 73019, USA; 34, Departments of Genetics and Medicine and Harvard-Partners Center for Genetics and Genomics, Harvard Medical School, Boston, Massachusetts 02115, USA; 35, Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 36, Department of Computer Science, University of California, Santa Barbara, California 93106, USA; 37, US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA; 38, Department of Computer Science, University of Western Ontario, London, Ontario N6A 5B7, Canada; 39, Cold Spring Harbor Laboratory, PO Box 100, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA; 40, Wellcome Trust, 183 Euston Road, London NW1 2BE, UK; 41, Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0114, USA; 42, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; 43, Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453, USA; 44, Bioinformatics Solutions Inc., 145 Columbia Street W, Waterloo, Ontario N2L 3L2, Canada; 45, Department of Molecular and Human Genetics, Baylor College of Medicine, Mailstop BCM226, Room 1419.01, One Baylor Plaza, Houston, Texas 77030, USA; 46, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02138, USA

* Members of the Mouse Genome Analysis Group

Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs

The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team*

*A full list of authors appears at the end of this paper

Only a small proportion of the mouse genome is transcribed into mature messenger RNA transcripts. There is an international collaborative effort to identify all full-length mRNA transcripts from the mouse, and to ensure that each is represented in a physical collection of clones. Here we report the manual annotation of 60,770 full-length mouse complementary DNA sequences. These are clustered into 33,409 'transcriptional units', contributing 90.1% of a newly established mouse transcriptome database. Of these transcriptional units, 4,258 are new protein-coding and 11,665 are new non-coding messages, indicating that non-coding RNA is a major component of the transcriptome. 41% of all transcriptional units showed evidence of alternative splicing. In protein-coding transcripts, 79% of splice variations altered the protein product. Whole-transcriptome analyses resulted in the identification of 2,431 sense-antisense pairs. The present work, completely supported by physical clones, provides the most comprehensive survey of a mammalian transcriptome so far, and is a valuable resource for functional genomics.

With the availability of draft sequences of the human genome^{1,2}, increasing attention has focused on identifying the complete set of mammalian genes, both protein-coding and non-protein-coding. As various analyses^{3,4} have noted, different annotation criteria lead to different sets of predicted genes, and even the true number of protein-coding genes remains uncertain. Biases in the training sets used for optimizing gene-finding algorithms lead to systematic biases in the types of genes that are predicted computationally—and so important genes and gene classes can be missed. A systematic analysis of transcripts from human chromosome 21 using dense oligonucleotide arrays revealed that the apparent transcriptional output of processed cytoplasmic messenger RNAs exceeds the predicted exons by an order of magnitude⁵.

One significant class of 'genes' missing from the existing genome annotation are those that give rise to non-protein-coding RNAs. Non-coding RNAs, although not highly transcribed⁶, constitute a major functional output of the genome. In addition to their role in protein synthesis (ribosomal and transfer RNAs), non-coding RNAs have been implicated in control processes such as genomic imprinting⁷ and perhaps more globally in control of genetic networks⁸.

Because of these and other limitations, it is clear that the utility of mammalian genome sequences will be more fully realized when there is a complete description of a mammalian transcriptome. The transcriptome includes all RNAs synthesized in an organism, including protein-coding, non-protein-coding, alternatively spliced, alternatively polyadenylated, alternatively initiated, sense, antisense, and RNA-edited transcripts. Attempts to catalogue the mammalian transcriptome have been based upon assembly of sequences from large-scale expressed-sequence-tag (EST) sequencing projects⁹. More recently, serial analysis of gene expression¹⁰ and large-scale sequencing of open reading frame (ORF) sequence tags¹¹ have contributed to the definition of the transcriptome and have improved genome annotation. Each has the limitation that no physical cDNA clones are fully sequenced. The RIKEN Mouse Gene Encyclopedia Project is a large-scale effort to isolate and sequence novel full-length mouse cDNAs. The initial results and validation of our approach have been reported previously, and resulted in the functional annotation of 21,076 full-length cDNAs (FANTOM1)¹². Here we describe the characterization and annotation of the FANTOM2 clone set (60,770), consisting of

39,694 new cDNA clones (FANTOM2 new set) and the 21,076 FANTOM1 clone set. We also present a global analysis of the mouse transcriptome based upon an integration of the FANTOM2 sequences with all available mouse mRNA data from the public sequence databases¹³. This data set provides the most comprehensive view so far of the transcriptional potential and diversity within the mouse genome, and serves as an important model for analysing the transcriptomes of other higher eukaryotes.

To enable a global description of the transcriptome, we use the term transcriptional unit (TU) to describe a segment of the genome from which transcripts are generated. A TU is defined by the identification of a cluster of transcripts that contains a common core of genetic information (in some cases, a protein-coding region). The main advantage of this approach is that the definition is purely computational and unequivocal. The existence of a TU is inferred from the identification of mRNAs via full-length cDNA isolation and sequencing. A single cDNA sequence may define a TU. When multiple cDNA transcripts that share DNA sequence are identified, we define these sequence-based clusters as a single TU. For each TU, the 5' boundary is defined at the most distal transcription start site, and the 3' boundary at the most extreme poly(A) sequence. TUs are DNA strand-specific, and are typically bounded by promoters at one end and termination sequences at the other. With this definition, alternative spliced transcripts, alternative 5' start and 3' ends, and variants arising from recombination (as in the immunoglobulin and T-cell receptor loci) are subsumed within a single TU, even though they may generate protein or RNA products with very different functions. TUs on opposite strands are counted separately, even if they overlap spatially in the genome. Thus, antisense transcripts are considered to be made from separate TUs.

The FANTOM2 clone set

The cDNA clones used in the project were selected from 246 full-length enriched cDNA libraries, most of which were also normalized and subtracted, with most from C57BL/6J mice, as described elsewhere^{12,14,15}. Information about tissue source and other information regarding these libraries is available in Supplementary Information section 1.

1,442,236 sequences were grouped into 171,144 3'-end clusters

based on sequence similarity (Supplementary Information section 2)^{14–16}. Of these, 159,789 had no significant BLAST hit to known mouse genes, and were considered potentially novel. However, from the annotation of the FANTOM1 clone set¹² it became clear that alternative polyadenylation is common in the mouse transcriptome, and that the set of 3'-end clusters is significantly redundant. To address this problem, 547,149 of these clones were sequenced from their 5' ends to provide additional discrimination.

Representative cDNAs were selected to represent target clusters and fully sequenced as described in Supplementary Information section 2. In total, we generated 60,770 high-quality, full-insert cDNA sequences with an average length of 1.97 kilobases (kb) (Supplementary Information section 3). If the smaller FANTOM1 clones are removed, the average insert size is 2.35 kb. The sequences of all of the cDNAs are available from DDBJ/EMBL/GenBank. A summary of all the sequence quality information, including the distribution of Phred/Phrap scores, is provided in Supplementary Information section 4. This table also demonstrates the excellent agreement between the cDNA sequences and the completed mouse genome sequence. Overall, there was 99.8% identity of aligned bases, and more than half the differences were mismatches rather than insertions or deletions. Among the 14,276 clones having >98% similarity to SWISS-PROT + TrEMBL cDNAs, automatic non-annotated alignment suggested that at least 67% of the cDNA clones were completely full length (Supplementary Information section 5) including alternative amino or carboxy terminals, while about 4% showed complete alignment but for unknown reasons lacked start or stop codons.

Annotating and mapping the RIKEN mouse cDNAs

To expedite manual annotation, the 60,770 cDNAs were first subjected to automated annotation. The decision tree used for automated annotation is presented in Fig. 1 and Supplementary Information section 6. Briefly, clone sequences that had a high degree of similarity to known genes in the Mouse Genome Informatics (MGI, <http://www.informatics.jax.org/mgihome/>) and LocusLink/RefSeq (<http://www.ncbi.nlm.nih.gov/LocusLink/>)¹⁷ databases were assigned the official gene name and available Gene Ontology¹⁸ (GO) terms. Sequences that were considered possible paralogues to known mouse genes or possible orthologues of genes from another mammalian species based upon an imperfect match to previously characterized genes were annotated using an analogous process—except that the name assigned to the clone sequence was qualified with the suffix 'homologue', or the prefix 'similar to' or 'weakly similar to', depending on the degree of similarity of the protein match. Unassigned sequences with matches to mouse mRNA/EST clusters (UniGene or TIGR's tentative consensus sequences) were assigned the annotation of the EST cluster. Clones with good coding sequence predictions in which the only available information was a match to a known protein domain or predicted motifs were assigned a name based on the domain match. Clone sequences with no nucleotide or protein match but with a valid coding sequence prediction of at least 100 amino acids in length were called 'hypothetical proteins'. Clones for which there was an independent match to at least one other mouse EST from a distinct library were classified as 'unknown EST'. All other clones were labelled 'unclassifiable'. Throughout the process, an effort was made to use only informative names.

The results of the automated computational analyses were reviewed by an international consortium of sequence annotators and gene-family experts over 2 months during the Mouse Annotation Teleconference for RIKEN cDNA sequences (MATRICS). These annotators altered the computational annotation in 24.8% of cases. The best coding sequence (CDS) was chosen by MATRICS annotators using a standardized set of criteria. If no good CDS existed, annotators classified sequences as unknown, or alternatively as 5' untranslated region (UTR), or 3' UTR if there was evidence of

alignment with a longer transcript from a known gene. Clones with CDS of less than 100 amino acids were annotated only if supported by known proteins, motifs, predicted signal peptide, transmembrane regions or splicing evidence.

The FANTOM2 cDNA annotation system used in MATRICS also provides a tool for continued annotation and a high-quality viewer for each transcript (Supplementary Information section 7). The FANTOM2 annotation viewer is accessible at <http://fantom2.gsc.riken.go.jp/>. The results of the annotation for the representatives of these clones are described below.

The cDNA sequences were provided to the Mouse Genome Sequencing Consortium, and have been used in their genome sequence annotation in the accompanying report¹⁹. We independently mapped the cDNA sequences to the mouse genome sequence (MGSCv3) assembly (ftp://wolfram.wi.mit.edu/pub/mouse_contigs/MGSC_V3) as described in Methods. This provided a basis for determining orthology between mouse and human genes in the annotation process, based on knowledge of conserved linkage between these two species²⁰. Genome analysis also provided insight into the intron–exon structure of mouse genes. The unambiguously assigned chromosomal locations are summarized, and distribution of the clones along the mouse chromosomes is presented, in Supplementary Information section 8.

Analysis of the transcriptome

Creating a comprehensive transcript data set for the mouse

To enable an overview of the transcriptome, a single transcript sequence from each TU was chosen as a representative. Because the focus of the RIKEN Mouse Gene Encyclopedia Project has been to identify and sequence novel transcripts, the 60,770 FANTOM2 cDNA set does not contain a representative of all known mouse TUs. We therefore derived a comprehensive representative transcript and protein set (RTPS), using a layered strategy of redundancy detection to identify transcripts that correspond to the same TUs, first in the FANTOM2 collection and then among representative sequences of the FANTOM2 TUs and known mouse cDNAs

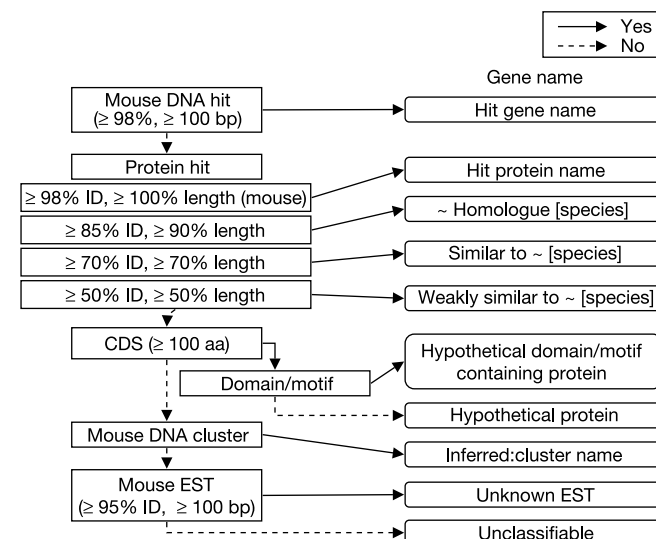


Figure 1 Flow chart of gene-name annotation pipeline. Nomenclature and criteria used in the gene name are described on the right-hand side. Gene names were annotated (right-hand side) from the gene function descriptor of the reference (left-hand side) according to the nomenclature. Priority was given to reference descriptors from which functional information could be inferred, even if references with less informative descriptors were more similar to the clones. CDS, coding sequence; ID, identity; aa, amino acid; bp, base pair; EST, expressed sequence tag.

(Supplementary Information section 9). The FANTOM2 cDNAs were first clustered by sequence similarity using ClusTrans (see Methods). Manual annotation corrected the number of clusters by ClusTrans from 35,957 to 35,637 and in some cases changed the representative clones. Of the 35,637 representative sequences, 11,108 were derived from multiple sequence clusters, while 24,529 were unclustered singletons.

The set of 35,637 cluster-resolved FANTOM2 representatives were then combined with a redundant set of 44,106 transcripts for mouse mRNA sequences compiled from LocusLink, MGI and GenBank (non-EST) databases (mouse public data set) (Supplementary Information section 9). The method for redundancy reduction and clustering criteria are outlined in Supplementary Information section 21. This step reduced the number of unique representative FANTOM2 sequences to 33,409 and the total number of representative transcripts in the RTPS to 37,086 (Fig. 2). 10,009 of the 33,409 TUs within the 60,770 FANTOM2 cDNA clone set correspond to known mouse genes of non-FANTOM2 sequences characterized in public databases (MGI, NCBI LocusLink, GenBank). The FANTOM2 data set represents an additional 15,923 novel TUs, a significant addition to the elucidation of the mouse transcriptome. Among these, 4,258 TUs contain protein-coding transcripts, and the remaining 11,665 TUs lack protein-coding potential. The proportion of novel non-protein-coding transcripts increased in the FANTOM2 new set compared to the FANTOM1 set. This could be a consequence of the intensive library normalization and subtraction strategy designed to collect less-abundant transcripts. The implication is that non-coding RNAs are relatively rare. The features of non-coding RNAs compared to the protein-coding RNAs are described later, and are shown in Supplementary Information section 14.

Functional annotation and CDS analysis by manual annotation

Table 1 represents the results of gene annotation for the FANTOM2 clone set. Of the 33,409 representatives, 13,736 (41.1%) were

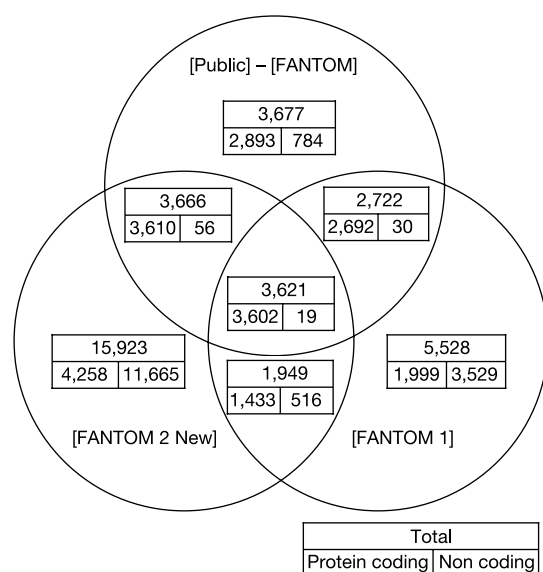


Figure 2 Breakdown of protein-coding and non-protein-coding transcriptional units (TUs) in three categories. [Public] consists of 44,106 sequences derived from the mouse public data set (transcripts for mouse mRNA sequences compiled from LocusLink, Mouse Genome Informatics and GenBank (non-EST) databases). [FANTOM1] contains 21,076 sequences that have been reported previously¹². [FANTOM2 New] contains 39,694 new sequences reported in the present work. In each box, the top row shows the count of TUs in each category, and the breakdown (protein-coding and non-protein-coding) is shown in the bottom row.

assigned some functional information. 6,929 (20.7%) TUs were assigned gene names from the categories of 'known mouse DNA', 'known mouse protein' or 'inferred from EST/mRNA cluster', which are the exact matches to a mouse gene with a known function. The categories 'homologue', 'similar to', 'weakly similar to' and 'domain/motif' represent functional information for 4,623 TUs (13.8%), inferred from sequence similarity. Others were assigned as 'hypothetical protein', 'unknown EST' or 'unclassifiable', hence no functional information was available. These TUs are potential candidates for novel genes.

Manually annotated CDS status was also evaluated for the representative clones of the FANTOM2 set (Supplementary Information section 10). Of the 33,409 FANTOM2 representative TUs, 17,594 were considered to have coding potential. Of these, 12,267 (69.7%) and 2,078 (11.8%) were translated into complete and partial protein sequences, respectively. Combining the protein sequences of FANTOM2 with those of public databases, the estimate of the protein coding TUs in the RTPS was 20,487.

Additional TU predicted by 5' and 3' EST mapping

In addition to mapping the cDNA sequences to the MGSCv3 genome to assist annotation, we also mapped 547,149 5' ESTs and 1,442,236 3' ESTs from the phase 1 sequencing and non-RIKEN contributions to a public EST database (NCBI dbEST) (Supplementary Information section 20). EST clusters were mapped to the mouse genome assembly based upon a minimal match of 95% over 100 base pairs (bp). The number of candidate TUs that were aligned outside the regions encompassed by the boundaries of TU within the RTPS is shown in Table 2. 5' end, 3' end and 5' + 3' pairs were counted separately. Of these, 3,425 candidate TUs that have both 5' + 3' pairs with additional EST support provide a minimal additional count to the 37,086 TUs in the RTPS. 3,422 5' ESTs and 6,703 3' ESTs are also supported by other ESTs. If we include singletons, the number of candidate TUs is estimated at 70,000.

Assignment of the transcription start sites of TUs

The correlation of expression measurements with identified promoter regions permits the association of specific *cis*-acting elements with particular patterns of expression²¹. In the mouse, this will only be possible when there is a complete enumeration of the transcription start sites for all TUs, so that the flanking sequences can be identified and analysed to identify promoters. To assess the reliability of 5' end assignment, we performed an alignment of FANTOM2 and RIKEN 5' ESTs to known genes. 131,335 sequences were aligned with previously known complete cDNAs. Of these, 116,660 5' ESTs or complete cDNAs could be aligned on the 5' UTR of the corresponding gene. 57,636 (4,431 genes) extended at least 10 nucleotides further than the recorded 5' ends, and probably represent a more accurate assignment of the transcription start point; 29,359 (3,068 genes) were within 10 bp of the 5' end of the mRNA and confirmed the starting point; and 29,665 (3,021 genes) were at least 10 nucleotides shorter than the published starting point although they still carry the first ATG. Some truncated

Table 1 Gene name assignment for FANTOM2 clones

Category	TU	Clone
Mouse DNA	5,772	14,753
Mouse protein	855	2,379
Homologue	2,604	7,508
Similar	1,114	2,578
Weakly similar	905	1,937
Domain/motif	2,184	5,356
Hypothetical protein	3,471	6,284
Mouse DNA cluster	302	462
EST match	7,005	9,485
Unclassifiable	9,197	10,028
Total	33,409	60,770

Number of TUs and clones in each gene name category described in Fig. 1.

	Supported	All
5' end	3,422	8,612
3' end	6,703	17,826
5' + 3' pair	3,425	6,654

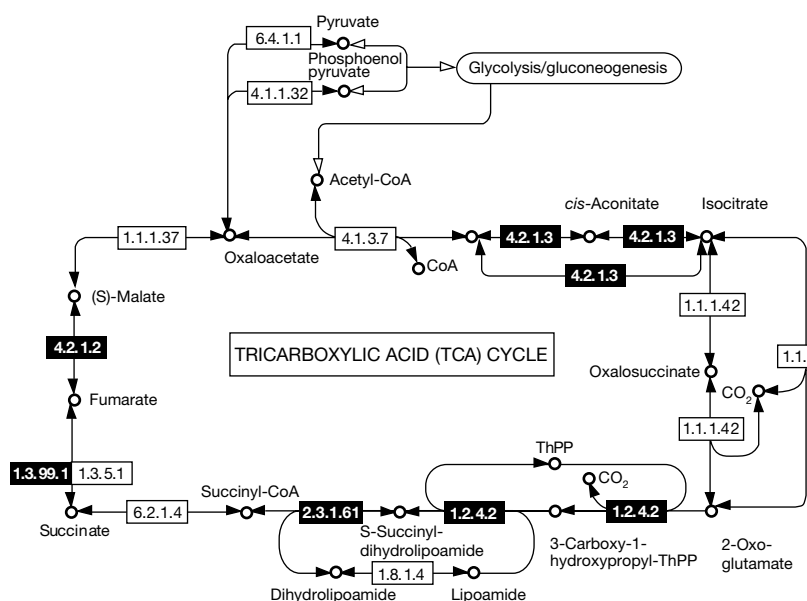
transcripts will represent alternative promoter use, or the known diversity of start sites in TATA-less promoters²².

The evolution of higher eukaryotes is associated with an increase in the number of introns and the incidence of alternative exon splicing. Previous EST-based estimates of alternative splice variants^{1,2,3} suggested that at least 59% of human and 33% of mouse genes are alternatively spliced. To assess the incidence of alternative splicing in mouse, we aligned 60,770 RIKEN clone sequences and the 44,106 mouse mRNA transcripts from the LocusLink, MGI (non-EST) and GenBank (non-EST) databases to the MGSCv3 mouse genome assembly using BLAST followed by SIM4 (ref. 24). 77,640 sequences out of 104,876 sequences were mapped to the mouse genome and were grouped into 33,042 clusters (Table 3). The removal of unspliced and single-exon derived transcripts, which are also irrelevant for splice analysis, reduced the number of clusters to 18,970. Of these, 7,293 clusters were singletons and excluded from further analysis. 4,750 (41%) clusters out of the remaining 11,677 clusters showed evidence of cryptic exons (exons included in only a subset of transcripts from a cluster) or exon length variation (Table 3). Modrek *et al.*²⁵ obtained similar numbers (42%) for human splice forms. The details of the status of each of the exon types from the 4,750 clusters are shown in Supplementary Information section 11. All variant and non-variant cluster transcripts can be explored at <http://genomes.rockefeller.edu/MouSDB/>.

We compared the entire clone set to the database of regulatory elements specific to mRNA UTRs (<http://bighost.area.ba.cnr.it/BIG/UTRHome/>) using PatSearch²⁶. The incidence of UTR-specific functional motifs (Supplementary Information section 12) provided additional functional annotation. The relevant annotation for each pattern was added only to those clones where the nature of the predicted protein and the specific location of the pattern allowed a very conservative assessment of its biological activity. This allowed confirmation of gene annotation in instances where the CDS was only partially sequenced. Additionally for clones annotated for the SECIS element (selenocysteine insertion sequence), it was possible to correctly define the CDS as it is known that in this case an in-frame UGA stop codon codes for selenocysteine.

Insertion events in CDS of transcripts may have direct impact on the protein coding function. The FANTOM2 clone set contains 105,861 repeats, including 14,929 low-complexity regions (Supplementary Information section 13A). 48.5% (51,372) of repeats were found in 64.8% (16,473) of FANTOM2 RTPS representatives covering 16.5% of the TU lengths (Supplementary Information section 13B). Among cDNAs with an annotated CDS, 14.2% (1,232) carry one or more repeats that overlap with the CDS. The majority contain repeats of the SINE (short interspersed nucleotide element), simple repeats, LINE (long interspersed nucleotide element) and LTR (long terminal repeat) classes. Most simple repeats did not cause frame-shifts or premature stop codons, whereas the dispersed elements commonly changed the CDS. For example, the insertion of a SINE B2 element into the CDS of a transcript A630030N07 (AK041699) similar to the intron-less ubiquitin-conjugating enzyme *Mhr6bn* will disrupt translation or produce a non-functional protein. The inserted B2 may affect the transcription of the neighbouring genes by acting as Pol II promoter²⁷. Sequencing of genomes and cDNA collections from strains other than C57BL/6J will distinguish inserted elements that are a source of functional polymorphism from those that represent species-specific mutagenic events.

The largest class of new transcripts in the FANTOM2 set lack an



the public data set and the FANTOM2 clone set; numbers in black boxes are newly found in the FANTOM2 clone set. CoA, coenzyme A; ThPP, thiamine pyrophosphate.

apparent protein-coding region; 15,815 TUs in the FANTOM2 set may represent functional non-coding RNAs. Supplementary Information section 14 compares sequence information for protein-coding and non-coding transcripts in the FANTOM2 set. The sequence quality does not distinguish the two classes, and does not explain the failure to annotate a CDS. Annotators had access to sequence quality information and also to CDS predictions that take account of possible frame shifts. The clearest difference is that 71% of the non-coding transcripts are unspliced/single exon compared to genomic sequence, whereas 18% of protein-coding transcripts are unspliced/single exon. The finding of 4,148 non-coding RNAs showing splice evidence is also a very interesting new feature. The frequency of polyadenylation signals does not differ significantly between the two classes of transcripts, suggesting that most non-coding transcripts are indeed the product of RNA Pol II-mediated transcription, processing and nuclear export.

To identify the strongest candidates as functional and novel non-coding RNAs, we further eliminated from the 15,815 potential non-coding RNAs any with similarity to known protein sequences identified using BLASTX (E -value $<10^{-5}$). We mapped the resulting 12,382 sequences to the MGSCv3, and selected clones that aligned at greater than 90% identity over more than 90% of their length. To further filter the data, we used exon predictions from GENSCAN analysis of the genomic sequence. Because GENSCAN may miss non-coding regions and UTRs of protein-coding transcripts, we considered sequences mapping on the same strand and within 10 kb of any predicted exon as protein-coding candidates and eliminated these from further consideration. The remaining 4,280 transcripts were considered the strongest non-coding RNA candidates.

Functional non-coding RNAs are likely to be distinguished by their reproducible expression and some measure of sequence conservation to mouse, human, rat EST or human genome. We assembled a line of evidence to support non-coding RNA annotation, including the number of mouse, rat and human EST hits, homology with human genome sequence and CpG islands ((C+G)-rich region) in genomic regions 5' upstream of each candidate. The nucleotide sequences of 1,194 (27.9%) transcripts out of 4,280 matched previously sequenced mouse ESTs (Table 4). Of these 4,280 representative transcripts from the RTPS, 252 transcripts matched rat ESTs (5.8%), and 111 transcripts matched human ESTs. 454 (10.6%) sequences mapped to both the mouse and human genome sequences, and we were able to identify CpG islands upstream of the transcription start site for 919 (21.5%) transcripts. Additionally, 1,150 (26.9%) of the functional non-coding RNA candidates were spliced (which is not significantly different from the overall frequency in the class) and 323 (7.5%) were also identified as candidate antisense transcripts.

Even applying the most conservative criteria, this suggests that several hundred mouse transcripts are conserved, regulated, non-coding mRNAs. Although the functional role played by these

transcripts remains to be studied in detail, the significant fraction of the transcriptional potential devoted to non-protein-coding genes suggests that they indeed have important biological and regulatory roles.

Antisense RNA

There is increasing recognition that the production of RNA transcripts from both strands in a genomic locus can produce coordinate regulation. For example, *Air*, the antisense of *Igf2r*, contributes to the paternal imprinting of three genes (*Igf2r*, *Scl22a2* and *Scl22a3*)²⁸. We focused on antisense transcripts using a slightly different strategy from that used to identify non-coding RNA. First, we aligned the sequences in the FANTOM2 set and public mRNA sequences with the mouse genome MGSCv3, and searched for pairs of sequences that mapped to the same genomic region in opposite orientations. We identified 2,431 pairs of sense-antisense transcripts overlapping in the exons of the sense gene by at least 20 bases. Of these, 1,573 pairs were not previously identified, and 316 of sense-antisense pairs contained the strongest non-coding RNA candidates described above.

Analysis of the predicted proteome

Our analysis of the proteome is based primarily on the previously described representative protein set (RPS), which contained protein sequences for each nucleotide where prediction was possible. We also created a variant-based proteome set (VPS) combining RPS and complete protein sequences representing splice variants not included in RPS. The VPS includes variant forms of known genes identified by sequencing of the FANTOM2 clones. The table of protein families for RPS and VPS is presented in Supplementary Information section 15. This table also includes the number of known members of a given family, including previously known human orthologues and those newly discovered. This analysis provides some indication of the extent to which the RPS represents the proteome. The overall appearance of the predicted mouse proteome is similar to that of the human (<http://www.ensembl.org>) deduced by genomic sequencing. Although there are fewer sequences in the current human set, the proportion covered by known domain families is the same, and the most abundant families have approximately the same number of members in both proteomes.

Protein domain/motif analysis and functional assignment

The constructed mouse proteome sequence set was analysed using a number of protein domain and motif databases including InterPro²⁹, Superfamily³⁰ and MDS³¹. An InterPro analysis gives a general functional overview of the sequence set and a functional overview of the transcriptome. InterPro domains were discovered in about 70% of the RPS, similar to other eukaryotic proteomes (<http://www.ebi.ac.uk/proteome>). Further analysis using the protein sets for human and mouse from the International Protein Index (IPI; Table 5) with the RPS suggests that the general domain composition is very similar for all three sets. The domain distribution in RPS and VPS is also very similar for top InterPro entries. The total number of protein sequences in the FANTOM2 set covered by all domain

Table 3 Analysis of alternative splicing

Data type	RIKEN-only		*RIKEN + mouse GB + MGI + LL	
	Sequences	Clusters	Sequences	Clusters
Mappings†	51,741	30,447	77,640	33,042
'Spliced' transcripts	31,593	16,541	54,490	18,970
Singletons	9,260	9,260	7,293	7,293
Multiple sequences	22,333	7,281	47,197	11,677
Alt. splicing variants	10,601	3,121	22,150	4,750
Non-variants	11,732	4,160	25,047	6,927

*GB + MGI + LL: mouse mRNA GenBank, Mouse Genome Informatics (non-EST divisions), and LocusLink.

†Mapping parameter: uninterrupted alignment, except for the initial and terminal nucleotides, with the mouse genome sequence in a single orientation with at least 95% of nucleotides identically mapped, and every exon aligned at 95% identity to the genome.

Table 4 Evidence supporting non-coding RNA annotation

Result of each optional analysis for 4,280 candidates	No. of hits
Mouse EST hit	1,200 (28.0%)
Human EST hit	111 (2.6%)
Rat EST hit	252 (5.8%)
Human genome homology (>50% ident., >70% length)	454 (10.6%)
Potential CpG islands	919 (21.5%)
Spliced sequences (no. of exons ≥ 2)	1,150 (26.9%)
Both in antisense and non-coding RNA candidates	323 (7.5%)

The number in parentheses indicates the rate of occurrence of non-coding RNA candidates among the extracted candidate transcripts.

databases (InterPro, Superfamily) is 17,236 (92% of the RPS). As expected, the most abundant domains are the zinc fingers, which are often duplicated in the same sequence; protein kinases are also highly abundant.

Superfamily domain analysis

Superfamily³⁰ analysis (http://supfam.org/SUPERERFAMILY/cgi-bin/gen_list.cgi?genome=mr) is used to detect and classify evolutionarily related groups of domains for which there is a known structural representative. An accurate superfamily definition is obtained by detailed manual analysis of structural, sequence and functional evidence for a common evolutionary ancestor³². The ancestral domain from each superfamily represents a genetic building block. These building blocks have been duplicated, recombined and mutated to create the proteins that are currently observed in the genome. A small number of domains have been duplicated a very large number of times, but most have been duplicated only a very few times. 98% of the domains identified by this analysis have been produced by duplication from 715 ancestral domains. The domain architecture for each sequence indicates the recombination of ancestral domains that has taken place during evolution.

Using strict criteria, a number of novel structural domain combinations were found in the FANTOM2 set. We identified 120 unknown, structural, pair-wise combinations of known structural domains; of these pairs, 30 have not been found previously in any sequenced proteome. As well as representing unique recombination events in evolution, these domain pairs provide targets for structural genomics projects that are assured to be novel. These newly discovered domain combinations can be found at <http://supfam.org/FANTOM2/domcombs.html>.

New domains

The availability of a very large set of new protein-coding transcripts, combined with existing information, permits computational detection of new motifs that are present in multiple independent gene products³¹. A number of new protein motifs, summarized in the MDS database (<http://motif.ics.es.osaka-u.ac.jp/FANTOM2/>), were predicted from the FANTOM2 sequences. One example is a structural GTPase submotif (MDS00154), specific for the immune associated nucleotide family (IAN), which also contain the canonical D-X-X-G pattern of the Walker B motif³³. The IAN specificity of the MDS00154 motif suggests that it modulates GTPase activity and specificity for signalling in the T-cell differentiation and selection process. For example, mouse *Ian1* was reported to be upregulated during the positive and β selection of thymocytes expressing TCR- β and TCR- α/β chains, a process that is essential to the development of a peripheral immune response³⁴.

Table 5 Top 10 InterPro entries

InterPro	Proteins matched (proteome coverage)			
	FANTOM2 RTPS	FANTOM2 VPS	IPI (<i>Mm</i>)	IPI (<i>Hs</i>)
IPR000694	988 (5.3%)	1,670 (4.9%)	528 (0.8%)	1,954 (1.6%)
IPR000822	453 (2.4%)	794 (2.4%)	1,017 (1.6%)	1,060 (0.9%)
IPR000719	417 (2.2%)	824 (2.4%)	299 (0.5%)	489 (0.4%)
IPR002290	398 (2.1%)	773 (2.3%)	271 (0.4%)	418 (0.3%)
IPR001245	382 (2.0%)	735 (2.2%)	254 (0.4%)	378 (0.3%)
IPR000276	353 (1.9%)	540 (1.6%)	205 (0.3%)	787 (0.6%)
IPR003599	309 (1.7%)	815 (2.4%)	280 (0.4%)	509 (0.4%)
IPR003006	280 (1.5%)	825 (2.4%)	381 (0.6%)	598 (0.5%)
IPR001356	212 (1.1%)	341 (1.0%)	40 (0.1%)	181 (0.1%)
IPR001680	209 (1.1%)	446 (1.3%)	144 (0.2%)	221 (0.2%)

IPR000694, proline-rich region; IPR000822, zinc finger, C2H2 type; IPR000719, eukaryotic protein kinase; IPR002290, serine/threonine protein kinase; IPR001245, tyrosine protein kinase; IPR000276, rhodopsin-like GPCR superfamily; IPR003599, immunoglobulin subtype; IPR003006, immunoglobulin/major histocompatibility complex; IPR001356, homeobox; IPR001680, G-protein β WD-40 repeat. RTPS, representative transcript and protein set; VPS, variant-based proteome set; IPI, International Protein Index; *Mm*, *Mus musculus*; *Hs*, *Homo sapiens*.

Membrane and secreted proteins

Membrane and secreted proteins are of particular interest because of their central role in intercellular communication and their accessibility as drug targets. We analysed the RTPS, using two independent methods to predict endoplasmic reticulum signal peptides³⁵ and to predict transmembrane helices, TMHMM 2.0 (ref. 36) and SVMtm (<http://genet.imb.uq.edu.au/predictors/>). First, we removed from the RTPS 1,559 readily identifiable partial ORFs that did not contain an initial methionine. As the two transmembrane helix prediction methods are subject to the misprediction of signal peptides as transmembrane domains, we developed a filter for predicted N-terminal transmembrane segments. If the predicted transmembrane's starting point was within the first 15 residues of the ORF and a signal peptide was predicted, then the region was regarded as a signal peptide rather than a transmembrane domain.

This analysis identified six classes of proteins on the basis of their membrane organization: (A) non-secretory proteins, (B) soluble/secreted proteins, (C) type I membrane proteins, (D) type II membrane proteins, (E) multi-span membrane proteins, and (F) unclassified proteins (Table 6). For each class, we adopted a stringent consensus method to restrict false-positive predictions, keeping only those motifs that were predicted using multiple methods. On the basis of this consensus prediction protocol, 80.1% (15,174) of the protein ORFs within the RTPS fall into classes A–E. Within the RTPS CDSs, 10.9% (1,877) were annotated as putative secreted proteins. Included among this class are the majority of soluble proteins involved in intercellular communication and the maintenance of the extracellular matrix, and 521 novel putative secreted proteins originating from the FANTOM2 project.

Small proteins

Functional proteins less than 100 amino acids in length were annotated only if they showed significant homology to known proteins from other species or members of gene families. On the basis of these stringent criteria, only 376 proteins of less than 100 amino acids were annotated. 4,558 contained predicted CDS regions encoding proteins between 50 to 99 amino acids in length, and these were reanalysed to identify high-quality, putative short CDSs. First, clones were eliminated if the CDS was not initiated within the first 500 nucleotides, on the assumption that few genuine 5' UTRs exceed this length. For each of the remaining 3,159 transcripts, we used tBLASTn to search the translated Ensembl human gene predictions. Among the 1,823 transcripts that had homology to a single entry in the human genome annotation, we searched for the potential splicing and identified 557 spliced transcripts. These are regarded as the most likely short protein candidates, although the possibility that some short CDSs are encoded by single exon TUs cannot be discounted. This analysis suggests that short proteins might increase the predicted proteome by as much as 10%, but each candidate will require further individual annotation and validation.

Table 6 Classification of 17,209 RTPS CDS proteins by membrane organization

Class	Signal peptide NN/HMM	Transmembrane TMHMM/SVMtm	RIKEN (%)	RTPS (%)
A	No/no	No/no	4,391 (63.2)	10,138 (58.9)
B	Yes/yes	No/no	521 (7.5)	1,877 (10.9)
C	Yes/yes	Yes/yes (single span)	188 (2.7)	734 (4.3)
D	No/no	Yes/yes (single span)	217 (3.1)	503 (2.9)
E	Any	Yes/yes (multi span)	667 (9.6)	1,922 (11.2)
F	No/yes or yes/no or no/no	No/yes or yes/no or no/no	964 (13.9)	2,035 (11.8)

Proteins are classified into six classes: A, non-secretory proteins; B, soluble/secreted proteins; C, type I membrane proteins; D, type II membrane proteins; E, multi-spanning membrane proteins; F, all the proteins not assigned by all above criteria. Neural Networks (NN) and hidden Markov model (HMM) are prediction methods of signal peptide. TMHMM2.0 and SVMtm (<http://genet.imb.uq.edu.au/predictors/>) are prediction methods for transmembrane helices.

Functional analysis of predicted proteins using Gene Ontology

To summarize the functional capacity of the transcriptome, we used the structured vocabularies of the Gene Ontology (GO) project³⁷ (<http://www.geneontology.org/>), as described in Methods. Of the 18,768 representative proteins, we were able to assign molecular function GO terms to 11,125 TUs, biological process GO terms to 10,443 TUs and cellular component GO terms to 10,488 TUs, representing 59.3%, 55.6% and 55.9%, respectively. GO annotations for individual genes are displayed in the FANTOM web interface, along with codes denoting the method used in the assignment (<http://fantom2.gsc.riken.go.jp/>). Supplementary Information section 16 summarizes the results of our analysis.

We also compared the results of our analysis with a similar binning of the Ensembl annotation of the human genome (Supplementary Information section 16). The distributions are very similar, re-emphasizing the importance of the mouse as a model system for human biology. Notable differences are seen in the areas where annotators focused on genes with special biological properties. For example, membrane proteins and extracellular proteins show significant increases in the mouse genome compared to the human genome: 2,892 versus 940 genes, and 4,907 versus 4,068 genes, respectively. Another area where the number of mouse genes was significantly greater than human genes, 3,152 versus 2,310, was in the molecular function 'transporter' and the biological process transport. The richness of the mouse annotations from this analysis provides an important foundation for the functional annotation and investigation of human orthologues.

Human disease genes

The catalogue of the mouse transcriptome can also be used to identify candidate genes associated with human diseases and mouse phenotypic variants. We obtained a list of 1,712 human loci associated with diseases by searching LocusLink (<http://www.ncbi.nlm.nih.gov/LocusLink>). Of these, 1,022 had associated sequence information, while 690 did not. We searched the 1,022 disease-associated protein sequences against the FANTOM2 data set using tBLASTn with a minimum *E*-value of less than 10^{-50} as the criterion for significance. Of 921 human diseases that had previously been identified with a homologous mouse gene, 740 (80%) were found in the FANTOM2 data set. For the 101 human disease genes without previously identified mouse homologues, 67 (66%) were mapped to the FANTOM2 data set (http://fantom2.gsc.riken.go.jp/supplement/disease_genes/).

Other gene discoveries

The FANTOM2 cDNA collection also revealed a substantial number of new protein-coding transcripts for which the likely function can be inferred from domain assignments or sequence homology (Supplementary Information section 17). For many of these proteins, we can also ascertain tissue-specific expression patterns from both the library of origin and the RIKEN Expression Array Database (<http://read.gsc.riken.go.jp>). Furthermore, using GO assignments and sequence similarity, we can assign these to gene families representing the full spectrum of functions of an archetypal mammalian cell. A number of highlights are described below.

- **Cell movement.** Miki *et al.*³⁸ have reported the identification of kinesin superfamily proteins (KIFs) in the mouse and human genome. KIFs are motor proteins contributing to intracellular trafficking by transporting vesicles, protein complexes and chromosomes³⁹. In FANTOM2, we found 33 KIFs equivalent to 73.3% of the 45 known loci. Of the 33 KIFs, 17 were full length. Previously, 2 alternative splice variants had been reported. In FANTOM2, we found 4 additional splice variants. Clones of 25 loci were derived from neural tissue or mixtures of neural and other tissue, consistent with previous reports.

- **Protein metabolism and turnover.** Ubiquitination and targeted protein degradation within the proteasome complex controls many

processes in eukaryotic cells, including meiosis, cellular proliferation and development⁴⁰. Unlike cell death and the cell cycle, which have been studied extensively, many ubiquitination regulatory proteins remain to be discovered in mammals. Our analysis identified a number of new gene products participating in this process, including 4 E1 ubiquitin activating enzymes, 13 E2 ubiquitin conjugating enzymes, 98 E3 ubiquitin ligases and 6 de-ubiquitinating enzymes. The extent of these gene families suggests that the targeting of cellular proteins for degradation is, indeed, a very highly regulated process.

- **G-protein-coupled receptors (GPCRs).** These comprise the largest family of receptor proteins in mammals, and may be promising drug targets. At present, there are approximately 600 intact GPCR genes that have been identified within the human genome, excluding nearly 350 odorant receptor genes^{1,2}. 410 clones in the FANTOM2 collection encoded candidate GPCRs; clustering to the TUs reduced this number to 213, although the high degree of relatedness and frequency of alternative splicing among some GPCR families make such clustering challenging. Out of these, 165 genes were contained within the 308 GPCRs previously annotated in the MGI database. The remaining 48 were unique to the mouse, and 14 of these had no clear mammalian orthologue. Among these, we identified two new members C630030A14 (AK083234) and 5330439C02 (AK030625) of G-protein-coupled receptor family C, which contains metabotropic glutamate receptors, γ -aminobutyric acid type B receptor, and Ca^{2+} -sensing receptors. These are highly homologous to previously identified GPCRs in the *Caenorhabditis elegans* and *Drosophila* genomes.

- **The 'metabolome'.** The InterPro and GO functional assignments allowed us to assign corresponding Enzyme Commission (EC) numbers⁴¹ to many of the predicted proteins with inferred metabolic enzymatic activities. In total, 726 distinct EC numbers were assigned to 3,583 TUs in the RTPS (approximately 10% of the total), of which nearly 90% (3,182) contained FANTOM cDNA clones. In comparison, the human enzyme class in the KEGG metabolic pathway database⁴² contains 720 unique EC number assignments.

The RTPS contains representatives of the mouse orthologues of most genes involved in known metabolic pathways. For example, the entirety of the tricarboxylic acid (TCA) cycle (Fig. 3) appears in the RTPS. All the enzymes in the TCA cycle were covered by FANTOM2 clones.

- **The impact of alternative splicing on the proteome.** As alternative splicing can change protein function, we evaluated its effect on putative translations of 4,750 variant clusters consisting of 22,150 sequences (Table 3), which showed evidence of cryptic splicing or exon length variation. Of the 4,750 variant clusters, 4,263 clusters were potentially protein coding. The analyses of all the exons derived from these 4,750 variant clusters are shown in Supplementary Information section 11. 8,500 exons (2,530 constitutive exons with lengths variation + 3,338 cryptic internal exons + 2,582 cryptic terminal exons) exhibited length variation. Of these 8,500 variant exons, 6,247 (73.5%) are within the CDS and will affect the CDS of 3,378 (79.2%) variant clusters out of 4,263 potential protein-coding clusters. The high number of splice forms translated into potential variant proteins is in line with a previous estimate that 74% of splice variation in human transcripts alters the coding sequence²⁵.

Alternative splicing, especially exon-skipping, can radically change the function of the protein product, and many examples can be seen in our data set. For example, the tyrosine phosphatase receptor type R (D130067J21 (AK051719)) of variant cluster sc1364 (1500019M22 (AK028054), D130067J21 (AK051719) and RefSeq NM_011217) lacks a transmembrane domain owing to alternative splicing. However, the sequence retained a cleavable signal peptide, implying that it could be secreted from cells. The variant also lacks the tyrosine phosphatase domain, and may therefore act as a dominant-negative form of the full-length recep-

tor. Other splice variants among putative signalling proteins cause loss of protein interaction domains. Examples include PDZ (Rap guanine nucleotide exchange factor, scl1714), pleckstrin (Rho interacting protein 3, scl1734), phosphatase (putative phosphatase, scl1827) and leucine-rich repeat (S-phase kinase-associated protein 2, scl2706), whose splice variants may generate proteins that have very different functions in cellular regulation than do their canonical forms.

Discussion

The transcriptome discovery strategy

Our aggressive approach to cDNA library normalization and subtraction, and the use of cDNA libraries from a wide range of tissues, has succeeded in providing the largest set of independent and distinct full-length cDNAs available for any species. In the early stages of this project, smaller clones were prioritized for full-length sequencing¹². As the project progressed, larger cDNAs were selected to facilitate completion of the transcriptome. This places greater demands on the full-length cDNA synthesis process, and may potentially result in truncated transcripts. The successful application of library construction strategies developed by the RIKEN team has allowed a high success rate in identification of full-length transcripts. This is evidenced by our identification of many new cDNAs well over 4 kb in length, including those encoding G-protein-coupled receptors and ion channels whose sequences have been validated by mapping to the draft mouse genome.

One consequence of the selection for both rare transcripts and for longer cDNAs is that the largest class of new and novel cDNAs are non-coding. Our analysis shows that the majority (~80%) of non-coding transcripts are not spliced. Sequence alone cannot distinguish genuine functional unspliced non-coding RNAs (such as *Air*²⁸) from unspliced heterogeneous nuclear RNAs (hnRNAs) that might be included in libraries through the presence of an internal poly(A) sequence. The rigorous isolation of cytoplasmic mRNA⁴³ introduced in later stages in this project may reduce the incidence of hnRNA if it does make a significant contribution. However, a disadvantage of targeting cytoplasmic RNAs is the assumption that all functional RNAs enter the cytosol. Such a strategy might also exclude non-coding RNAs that are active in the nucleus and that may have a regulatory role by interfering with splicing, polyadenylation or nuclear export. A large proportion of the non-coding transcripts are supported by independent ESTs. By definition, the 'unclassifiable' class represents singletons, but around 60% of 'unclassifiable' transcripts analysed thus far are detected reproducibly using cDNA microarrays (<http://read.gsc.riken.go.jp>). This finding supports the analysis of polyadenylation, which indicates that the majority are genuine mature products of transcription mediated by pol II. Overall, it is clear that non-coding RNAs are a major component of the mammalian transcriptome.

The size of the transcriptome and the proteome

Much of the discussion arising from completion of mouse and human genome sequences has revolved around estimating the number of genes. To analyse the draft transcriptome, we developed a stringent definition of a TU as a unit of genetic information transcribed into mRNA. Our analysis of the available sequence data from this project and of that previously reported in the public databases suggest that the mouse genome contains 37,086 TUs. However, it should be noted that even this is an incomplete survey of the transcriptional potential in mouse. For example, of the 22,444 protein-coding genes predicted with a high level of confidence in mouse Ensembl¹⁹, 3,074 were not covered by RTPS or by any RIKEN 5' or 3' EST sequencing.

The number of clusters is, of course, a function of the clustering parameters. Many larger clusters, some containing up to 90 representatives, may contain multiple representatives of gene families.

Examples include the *Gapd*, *RpL21*, *RpL29*, *Hmgb1*, *Rps2* and *RpL7a* clusters. The sequences in each of these clusters are very similar over their entire length, and readily meet stringent clustering cut-offs, but they clearly derive from different loci and should therefore be considered separate TUs. For example, *Gapd* is known to have 70 loci in the mouse genome (http://www.ncbi.nih.gov/LocusLink/list.cgi?V=1&Q=NM_008084).

As noted previously, there are a large number of 3' and 5' EST clusters from the phase 1 sequencing project that remain to be sequenced. To make a better estimate of the number of TUs in the mouse genome, we analysed the 547,149 5' ESTs and 1,442,236 3' ESTs from the phase 1 sequencing and non-RIKEN entries in dbEST. These ESTs were mapped to the draft mouse genome sequence. To count the number of additional TU candidates, we counted the number of EST clusters that were mapped outside the region of TUs determined by RTPS (Table 2), thereby increasing the estimated number of TUs in the mouse genome to around 70,000.

In addition, we have a number of reasons to suspect that our present estimate represents a conservative lower bound to the number of TUs in the mouse genome:

(1) Members of the FANTOM consortium and the RIKEN Genome Exploration Research Group continue to collaborate in the production of full-length cDNA libraries from novel tissue sources. In these continuing efforts, which use a very aggressive approach to normalization and subtraction of cDNA libraries, novel TUs are being discovered at a significant rate. Furthermore, much of our effort is now focused on inducible genes expressed in cell types such as macrophages, lymphocytes, dendritic cells, melanocytes and specialized cells in the central and peripheral nervous systems.

(2) Long mRNAs on the scale of genes such as *titin*⁴⁴ and *dystrophin*⁴⁵ are under-represented in all available cDNA libraries owing to the limitations in reverse transcription, cloning and stability of long cDNA clones; such genes are also very difficult to identify and annotate in genomic sequence. Continuing efforts are directed at selected cloning of longer mRNAs using novel RIKEN full-length cloning strategies⁴⁶.

(3) Our definition of TUs did not account for the fact that some overlapping transcripts may encode distinct functions and may be independently regulated. Complete functional annotation of the genome will require a development of criteria that will allow such clusters to be separated. Transcript 6030402A12 (AK031286), which contributes a single TU to our estimate, contains at least 12 distinct transcripts that differ at both the 5' and 3' ends and consequently include quite different protein functional domains. Some of this variation has been noted previously⁴⁷. By alignment to the mouse genome sequence, we can demonstrate that proteins that contain both N-terminal activation (LER) and repression (KRAB) domains—or proteins that lack either of these—arise from alternative promoter usage. Alternative polyadenylation (as well as internal exon skipping) yields C-terminal C2H2 zinc-finger arrays containing zero, 5, 6, 7, 8, 9, 17, 18 or 21 repeats in individual variant forms. It is not self-evident that the SCAN-KRAB zinc-finger locus contains a single TU (as defined here), or a single gene.

We can begin to infer that estimates of the number of TUs in the human genome based on EST assembly⁹ were not unrealistic. At least 41% of TUs encode more than one form of mRNA, consistent with computational analysis in the Human Alternative Splicing Database²⁵ (<http://www.bioinformatics.ucla.edu/~splice/HASDB/>). This will certainly be an underestimate, because our approach sought to minimize the number of replicate sequences from a single TU. Furthermore, tissue-specific or infrequently spliced variants are sampled less frequently by the cloning strategy that we used, and might be eliminated through subtraction or clone selection. Alternative splicing cDNA libraries⁴⁸ and oligonucleotide-based microarrays⁴⁹ could be used to expand our knowledge of the impact of alternative splicing in the mouse transcriptome and proteome. Depending on how alternative 5' ends, 3' ends and cryptic exons are

combined, the total number of transcripts will be at least twice the number of TUs.

Final comments

With the simultaneous, and synergistic, availability of genome and transcriptome sequences, we have reached a milestone in the era of large-scale data acquisition. We expect that additional mammalian transcriptome projects will be guided by the experience of the Mouse Gene Encyclopedia project, and that their outputs will support and refine the functional annotation of the TUs described here. As we continue this project, we hope that we can provide the tools necessary for a complete understanding of the biological processes that underlie mammalian development, growth, adaptation and disease. On the basis of the analysis presented here, we believe that a genome sequence alone can provide us with, at best, an incomplete picture of the transcriptional capacity of a complex genome. Looking forward, we anticipate that transcriptional analysis, both through sequencing and large-scale expression analysis, will be essential to the building of a complete picture of how genes and their products interact to form the regulatory, signalling and metabolic pathways that are the basis for complex life on Earth. Finally, although the mouse may be an imperfect model for the understanding of human biology, we believe that the resources made available through our efforts and those of the Mouse Genome Sequence Consortium—in combination with the exquisite genetic resources that are already available—will make the mouse the most appropriate animal for human studies for years to come. □

Methods

Creation of the cDNA resource

Library construction, clone selection and sequencing is described in detail in Supplementary Information section 2. The cDNA sequence was assembled with the Phred/Phrap/Consed system, taking advantage of the anchor sequences of the 5' and 3' ends of cDNAs. Sequence and quality data for all assembled clones is available through DDBJ/EMBL/GenBank (Supplementary Information section 18), as well as through the FANTOM website (<http://fantom2.gsc.riken.go.jp>).

Mapping sequences to mouse draft genome

Sequences to be mapped were first screened using RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>). We aligned the full-length sequences with the draft mouse genome sequence (ftp://wolfram.wi.mit.edu/pub/mouse_contigs/MGSC_V3), using Parcel BLAST, with an *E*-value of 10^{-5} . We extracted the genomic region of the best locus assigned to each cDNA (>95% identity over >100 bases), and used SIM4 (ref. 24) to align the cDNA to the genomic region. We aligned the repeat-masked EST sequences with the mouse genome, using BLAT⁵⁰.

Clustering of FANTOM2 clone set and RTPS using ClusTrans

The 60,770 FANTOM2 clone set was first clustered using our original cDNA clustering program ClusTrans. Pairwise comparisons and global alignment were performed for all cDNAs using the SSEARCH program distributed with FASTA^{51,52}, using the following parameters: match score, +1; mismatch score, -2; penalty for the first residue in a gap, -8; and penalty for additional residues in a gap, 0. After the global alignment, cDNA clusters were defined on the basis of percentage sequence identity and match length. The detailed criteria for defining a cluster are summarized in Supplementary Information section 19.

Creating an RTPS

To derive a comprehensive yet unique representative transcript and protein set (RTPS), the 60,770 FANTOM2 cDNA sequences, clustered initially with ClusTrans as above, were further clustered with 44,106 mouse transcript sequences compiled from the LocusLink, MGI (non-EST) and GenBank (non-EST) databases. The detailed strategy for construction of the RTPS is described in Supplementary Information sections 19 and 21.

Each cluster in the combined FANTOM2/public domain compilation defines a TU, and a single transcript sequence was chosen to represent each TU. For protein-coding TUs, a single polypeptide sequence was chosen to represent each TU with the following selection rules, longest SWISS-PROT record > longest RefSeq protein (NP_) record > longest CDS sequence. The number of protein sequences successfully translated and included in the RPS was 18,768. A variant-based proteome set (VPS) was prepared by identifying alternative transcription products that have different CDS sequences compared to the CDS sequences of cluster representatives. Each FANTOM2 clone with a unique CDS due to alternative transcript production, contributed to the VPS. The VPS contains all 18,768 protein sequences in the RPS plus 15,518 coding sequences.

Assignment of Gene Ontology terms

For genes represented in the MGI databases (<http://www.informatics.jax.org/>), we

adopted GO terms that were assigned by MGI annotators⁵³. In some cases, MATRICS annotators assigned GO terms. Terms were also assigned to TUs using translation tables of SWISS-PROT keywords, InterPro domains, and EC assignments (<ftp://ftp.geneontology.org/pub/go/external2go>) to GO. A preliminary SCOP (Structural Classification of Proteins; <http://scop.wehi.edu.au/scop/>) structural domain to GO translation table, which is continuously being refined, was created for this project (<http://www.supfam.org/SUPERFAMILY/GO/>). GO binning strategies are described in Supplementary Information section 21.

Analysis of alternative splice variants

To confirm the splice candidates in the 4,750 variant clusters, variant exons were compared to mouse mRNA sequences in the dbEST division of GenBank. dbEST sequences were mapped to the mouse genome as for the mapping of full-length cDNAs: the locus was identified using BLAT⁵⁰, and intron/exon boundaries were identified using SIM4. ESTs which mapped with at least 95% identity to the genome, with every exon mapped at 95% identity or less than 5 gaps or mismatches, were interrogated for the presence of the exon forms observed in the cDNA data.

The impact of variant splicing on the translated protein sequences was assessed by comparing all variant clusters to the protein sequences of the FANTOM2 RTPS. cDNA-to-genome alignments were used to determine the genomic coordinate of the CDS start and end.

Databases, tools and systems

The databases, tools and systems used for the analyses in this paper are described in Supplementary Information section 20.

Received 19 September; accepted 28 October 2002; doi:10.1038/nature01266.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
4. Daly, M. J. Estimating the human gene count. *Cell* **109**, 283–284 (2002).
5. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
6. Harrison, P. M. *et al.* Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280 (2002).
7. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* **2**, 21–32 (2001).
8. Mattick, J. S. & Gagen, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18**, 1611–1630 (2001).
9. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
10. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
11. Camargo, A. A. *et al.* The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl Acad. Sci. USA* **98**, 12103–12108 (2001).
12. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
13. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
14. Carninci, P. *et al.* Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000).
15. Carninci, P. *et al.* Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* **77**, 79–90 (2001).
16. Konno, H. *et al.* Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res.* **11**, 281–289 (2001).
17. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
18. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
19. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* (this issue).
20. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
21. Pilpel, Y., Sudarsanam, P. & Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **29**, 153–159 (2001).
22. Smale, S. T. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta* **1351**, 73–88 (1997).
23. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nature Genet.* **30**, 29–30 (2002).
24. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974 (1998).
25. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29**, 2850–2859 (2001).
26. Pesole, G., Liuni, S. & D'Souza, M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics* **16**, 439–450 (2000).
27. Ferrigno, O. *et al.* Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nature Genet.* **28**, 77–81 (2001).
28. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).
29. Apweiler, R. *et al.* InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145–1150 (2000).

30. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).
31. Kawaji, H. *et al.* Exploration of novel motifs derived from mouse cDNA sequences. *Genome Res.* **12**, 367–378 (2002).
32. Murzin, A. G. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394 (1996).
33. Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41–72 (2002).
34. Poirier, G. M. *et al.* Immune-associated nucleotide-1 (IAN-1) is a thymic selection marker and defines a novel gene family conserved in plants. *J. Immunol.* **163**, 4960–4969 (1999).
35. Nielsen, H. & Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 122–130 (1998).
36. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
37. The Gene Ontology Consortium Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
38. Miki, H., Setou, M., Kaneshiro, K. & Hirokawa, N. All kinesin superfamily protein, KIF, genes in mouse and human. *Proc. Natl Acad. Sci. USA* **98**, 7004–7011 (2001).
39. Hirokawa, N. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science* **279**, 519–526 (1998).
40. Weissman, A. M. Themes and variations on ubiquitylation. *Nature Rev. Mol. Cell Biol.* **2**, 169–178 (2001).
41. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
42. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
43. Carninci, P., Nakamura, M., Sato, K., Hayashizaki, Y. & Brownstein, M. J. Cytoplasmic RNA extraction from fresh and frozen mammalian tissues. *Biotechniques* **33**, 306–309 (2002).
44. Bang, M. L. *et al.* The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.* **89**, 1065–1072 (2001).
45. Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509–517 (1987).
46. Carninci, P., Shiraki, T., Mizuno, Y., Muramatsu, M. & Hayashizaki, Y. Extra-long first-strand cDNA synthesis. *Biotechniques* **32**, 984–985 (2002).
47. Dreyer, S. D., Zheng, Q., Zabel, B., Winterpacht, A. & Lee, B. Isolation, characterization, and mapping of a zinc finger gene, ZFP95, containing both a SCAN box and an alternatively spliced KRAB A domain. *Genomics* **62**, 119–122 (1999).
48. Schweighoffer, F. *et al.* Qualitative gene profiling: a novel tool in genomics and in pharmacogenomics that deciphers messenger RNA isoforms diversity. *Pharmacogenomics* **1**, 187–197 (2000).

49. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
50. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
51. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
52. Pearson, W. R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650 (1991).
53. Hill, D. P. *et al.* Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics* **74**, 121–128 (2001).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank the following (listed in alphabetical order) for discussion, encouragement and technical assistance: K. Abe, T. Akimura, T. Hanagaki, K. Hayashida, K. Hiramoto, T. Hiraoka, M. Honda, F. Hori, T. Huber, M. Izawa, H. Kato, M. Kohda, Y. Kojima, S. Koya, K. Koyama, C. Kurihara, J. Mashima, T. Matsuyama, M. Murata, K. Nishi, K. Nomura, R. Numazaki, M. Ohno, H. Saitou, C. Sakai, H. Sano, Y. Shibata, A. J. G. Simpson, Y. Sogabe, M. Tagami, A. Tagawa, F. Takahashi, S. Takaku, Y. Takeda, T. Tanaka, A. Tomaru, W. W. Wasserman, A. Watahiki, C. Wicking, H. Yamanishi, T. Yao, K. Yoshida. We especially thank A. Wada, M. Muramatsu, T. Ogawa, A. Kira, and all the members of RIKEN Yokohama Research Promotion Division for supporting and encouraging the project. We also thank the Laboratory for Genome Exploration Research Group for secretarial and technical assistance. This work was mainly supported by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese Government to Y.H., and by a Research and Development for Applying Advanced Computational Science and Technology (ACT) grant from the Japan Science and Technology Corporation to Y.H. and J.K.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.H. (e-mail: yoshihide@gsc.riken.go.jp). Details of methods, and accession numbers for all 60,770 cDNA clones, are available as Supplementary Information. Clone availability information regarding access to the FANTOM2 clones is available at <http://www.riken.go.jp/>.

Authors' contributions Y. Okazaki, M. Furuno, T. Kasukawa, C. Schönbach, R. Baldarelli, D. P. Hill, C. Bult, D. A. Hume, J. Quackenbush, L. M. Schriml, A. Kanapin and Y. Hayashizaki are core authorship members; P. Carninci and J. Kawai are team organizers.

FANTOM Consortium: Y. Okazaki^{1,2}, M. Furuno¹, T. Kasukawa^{1,3}, J. Adachi¹, H. Bono¹, S. Kondo¹, I. Nikaido^{1,4}, N. Osato¹, R. Saito^{1,5}, H. Suzuki¹, I. Yamanaka¹, H. Kiyosawa^{1,4}, K. Yagi¹, Y. Tomaru^{1,6}, Y. Hasegawa^{1,4}, A. Nogami^{1,4}, C. Schönbach⁷, T. Gojobori⁸, R. Baldarelli⁹, D. P. Hill⁹, C. Bult⁹, D. A. Hume¹⁰, J. Quackenbush¹¹, L. M. Schriml¹², A. Kanapin¹³, H. Matsuda¹⁴, S. Batalov¹⁵, K. W. Beisel¹⁶, J. A. Blake⁹, D. Bradt⁹, V. Brusic¹⁷, C. Chothia¹⁸, L. E. Corbani⁹, S. Cousins⁹, E. Dalla¹⁹, T. A. Dragani²⁰, C. F. Fletcher^{15,21}, A. Forrest¹⁰, K. S. Frazer^{9,22}, T. Gaasterland²³, M. Gariboldi²⁰, C. Gissi²⁴, A. Godzik²⁵, J. Gough¹⁸, S. Grimmond¹⁰, S. Gustincich²⁶, N. Hirokawa²⁷, I. J. Jackson²⁸, E. D. Jarvis²⁹, A. Kanai⁵, H. Kawaji^{3,14}, Y. Kawasaki³⁰, R. M. Kedzierski³⁰, B. L. King⁹, A. Konagaya⁷, I. V. Kurochkin⁷, Y. Lee¹¹, B. Lenhard³¹, P. A. Lyons³², D. R. Maglott¹², L. Maltais⁹, L. Marchionni¹⁹, L. McKenzie⁹, H. Miki²⁷, T. Nagashima⁷, K. Numata⁵, T. Okido⁸, W. J. Pavan³³, G. Perte¹¹, G. Pesole²⁴, N. Petrovsky³⁴, R. Pillai¹⁷, J. U. Pontius¹², D. Qi⁹, S. Ramachandran⁹, T. Ravasi¹⁰, J. C. Reed²⁵, D. J. Reed⁹, J. Reid¹⁹, B. Z. Ring³⁵, M. Ringwald⁹, A. Sandelin³¹, C. Schneider¹⁹, C. A. M. Semple²⁸, M. Setou²⁷, K. Shimada^{36,37}, R. Sultana¹¹, Y. Takenaka¹⁴, M. S. Taylor²⁸, R. D. Teasdale¹⁰, M. Tomita⁵, R. Verardo¹⁹, L. Wagner¹², C. Wahlestedt³¹, Y. Wang¹¹, Y. Watanabe^{36,37}, C. Wells¹⁰, L. G. Wilming³⁸, A. Wynshaw-Boris³⁹, M. Yanagisawa³⁰, I. Yang¹¹, L. Yang⁹, Z. Yuan¹⁰, M. Zavolan²³, Y. Zhu⁹ & A. Zimmer⁴⁰

RIKEN Genome Exploration Research Group Phase I Team: P. Carninci², N. Hayatsu¹, T. Hirozane-Kishikawa¹, H. Konno¹, M. Nakamura¹, N. Sakazume¹, K. Sato⁹, T. Shiraki¹ & K. Waki¹

RIKEN Genome Exploration Research Group Phase II Team: J. Kawai^{1,2}, K. Aizawa¹, T. Arakawa¹, S. Fukuda¹, A. Hara¹, W. Hashizume¹, K. Imotani¹, Y. Ishii¹, M. Itoh², I. Kagawa¹, A. Miyazaki¹, K. Sakai¹, D. Sasaki¹, K. Shibata², A. Shinagawa¹, A. Yasunishi¹ & M. Yoshino¹

Mouse Genome Sequencing Consortium: R. Waterston⁴¹, E. S. Lander⁴², J. Rogers³⁸ & E. Birney¹³

Scientific management: Y. Hayashizaki^{1,2,4,6}

Affiliations for authors: 1, Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; 2, Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Main Campus, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan; 3, NTT Software Corporation, 223-1 Yamashita-cho, Naka-ku, Yokohama, Kanagawa, 231-8554, Japan; 4, Division of Genomic Information Resources, Science of Biological Supramolecular Systems, Graduate School of Integrated Science, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan; 5, Institute for Advanced Biosciences, Keio Univ, 403-1 Tsuruoka-city, Yamagata, 997-0017, Japan; 6, Institute of Basic Medical Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8577, Japan; 7, Biomedical Knowledge Discovery Team, Bioinformatics Group,

RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; 8, Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan; 9, Mouse Genome Informatics Group, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA; 10, Institute for Molecule Bioscience and ARC Special Research Centre for Functional and Applied Genomics, University of Queensland, Queensland 4072, Australia; 11, The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 12, National Center for Biotechnology Information, NIH, Bldg 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; 13, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 14, Graduate School of Information Science and Technology, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan; 15, Genomics Institute of the Novartis Research Foundation (GNF), 10675 John Jay Hopkins Drive, San Diego, California 92121, USA; 16, Boys Town National Research Hospital, 555 North 30th Street, Omaha, Nebraska 68131, USA; 17, Laboratories for Information Technology, 21, Heng Mui Keng Terrace, Singapore, 119613, Singapore; 18, Structural Studies, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; 19, LNCIB, Functional Genomics, AREA Science Park, Padriciano, 99 Trieste, 34012, Italy; 20, Istituto Tumori Milano, Milano, 20133, Italy; 21, The Scripps Research Institute, 10550N. Torrey Pines Road, La Jolla, California 92037, USA; 22, The Zebrafish International Resource Center, University of Oregon, Eugene, Oregon 97403-5274, USA; 23, Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA; 24, Università di Milano, Milano, 20133, Italy; 25, The Burnham Institute, 10901N. Torrey Pines Road, La Jolla, California 92037, USA; 26, Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA; 27, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan; 28, MRC Human Genetics Unit, Crewe Road, Edinburgh, UK; 29, Duke University Medical Center, Department of Neurobiology, Box 3209 Durham, North Carolina, 27710, USA; 30, Howard Hughes Medical Institute, Department of Molecular Genetics, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA; 31, Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, 17177 Stockholm, Sweden; 32, JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, Addenbrookes Hospital Hills Road, Cambridge CB2 2XY, UK; 33, National Human Genome Research Institute, National Institutes of Health, 49/4A82 49 Convent Drive, MSC4472, Bethesda, Maryland 20892-4472, USA; 34, Autoimmunity Research Unit, The Canberra Hospital, Yamba Drive, Woden, ACT 2606, Australia; 35, Applied Genomics, Inc. 525 Del Rey Avenue, Sunnyvale, California 94085, USA; 36, Hirakata Ryoikuen, 2-1-1 Tsudahigashi, Hirakata, Osaka, 565-0874, Japan; 37, Institute for Advanced Medical Sciences, Hyogo College of Medicine, Mukogawa 1-1, Nishinomiya, Hyogo, 663-8501, Japan; 38, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 39, University of California, San Diego School of Medicine, Pediatrics/Medicine 9500 Gilman Drive, La Jolla, California 92093-0627 USA; 40, Department of Psychiatry, University of Bonn, Sigmund-Freud-Strasse 25, Bonn, 53105, Germany; 41, Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA; 42, Whitehead Institute/MIT Center for Genome Research, 320 Charles Street, Cambridge, Massachusetts 02141, USA

The mosaic structure of variation in the laboratory mouse genome

Claire M. Wade^{*,†}, Edward J. Kulbokas III^{*}, Andrew W. Kirby^{*}, Michael C. Zody^{*}, James C. Mullikin[‡], Eric S. Lander^{*}, Kerstin Lindblad-Toh^{*,§} & Mark J. Daly^{*,§}

^{*} Whitehead Institute for Biomedical Research and Whitehead/MIT Center for Genome Research, 9 Cambridge Center, Cambridge, Massachusetts 02139, USA

[†] School of Veterinary Science, The University of Queensland, Queensland 4072, Australia

[‡] The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1RQ, UK

[§] These authors contributed equally to this work

Most inbred laboratory mouse strains are known to have originated from a mixed but limited founder population in a few laboratories^{1,2}. However, the effect of this breeding history on patterns of genetic variation among these strains and the implications for their use are not well understood. Here we present an analysis of the fine structure of variation in the mouse genome, using single nucleotide polymorphisms (SNPs). When the recently assembled genome sequence from the C57BL/6J strain³ is aligned with sample sequence from other strains, we observe long segments of either extremely high (~40 SNPs per 10 kb) or extremely low (~0.5 SNPs per 10 kb) polymorphism rates. In all strain-to-strain comparisons examined, only one-third of the genome falls into long regions (averaging >1 Mb) of a high SNP rate, consistent with estimated divergence rates between *Mus musculus domesticus* and either *M. m. musculus* or *M. m. castaneus*. These data suggest that the genomes of these inbred strains are mosaics with the vast majority of segments derived from *domesticus* and *musculus* sources. These observations have important implications for the design and interpretation of positional cloning experiments.

Patterns of genetic variation provide insight into the evolutionary history of a species and define the complexity of mapping phenotypes in that organism. The commonly used inbred laboratory strains of mice constitute the primary mammalian model system and are an integral component of medical genetic research. These inbred laboratory strains were predominantly derived in the early twentieth century from mouse breeders who originally bred 'fancy'

mice (for unusual coat colours and behaviours) as a hobby. Many of the most commonly used strains trace their origins to W. Castle's laboratory at Harvard University and even more strains originate from his supplier A. Lathrop of Granby, Massachusetts. Although these mice are generally thought to reflect predominantly the *M. m. domesticus* subspecies, there are some historical contributions from 'fancy' mice bred in Japan and China^{1,2} (Fig. 1a). As a result, we would expect to see in these strains recognizable contributions from several other subspecies such as *M. m. musculus* (and possibly *M. m. castaneus* through the hybrid *M. m. molossinus*). Indeed, most of these inbred laboratory strains carry a *M. m. musculus* Y chromosome⁴ (previous work had shown that most carry *M. m. domesticus* mitochondrial DNA^{5,6}).

The genomes of these strains were predicted to be a 'mosaic' of regions with origins in the different subspecies⁷, but a clear description of this variation has remained elusive, largely owing to a lack of high-resolution data across the genome. Here, we describe the fine structure of variation among inbred laboratory strains, first through an analysis of available finished sequence and then through more comprehensive analysis of genome-wide shotgun SNP discovery data produced by the Mouse Genome Sequencing Consortium³.

To study the nature of genetic variation among strains, we wished to compare the recent C57BL/6J (henceforth B6) draft genome assembly, MGSCv3 (ref. 3), with available finished sequence from other strains. We first, however, performed a control comparison of 205 finished B6 sequences (obtained from GenBank) totalling 49 megabases (Mb) with the draft genome assembly. Sequences were compared using SSAHA-SNP software⁸ to identify differences in high-quality matches. We detected 0.9 candidate SNPs per 10,000 base pairs (bp). The occurrence of candidate SNPs in this intra-strain (B6-B6) comparison suggests that these SNPs result from sequencing errors (rather than an unexpectedly high mutation rate within an ostensibly inbred strain). To clarify this, we re-sequenced 15 such candidate SNPs in mice from seven distinct B6 founder lines from the Jackson Laboratories as well as the DNA stock used in creating the draft genome sequence at the Whitehead Institute. None of the 15 apparent polymorphisms was confirmed—indicating that they are indeed sequencing artefacts. All 15 cases seem to result from errors in the MGSCv3 sequence, indicating a low error rate (0.9 per 10 kilobases, kb) that is actually below the target accuracy rate for 'finished sequence' of less than one error per 10 kb.

We then set out to perform the same comparison with finished

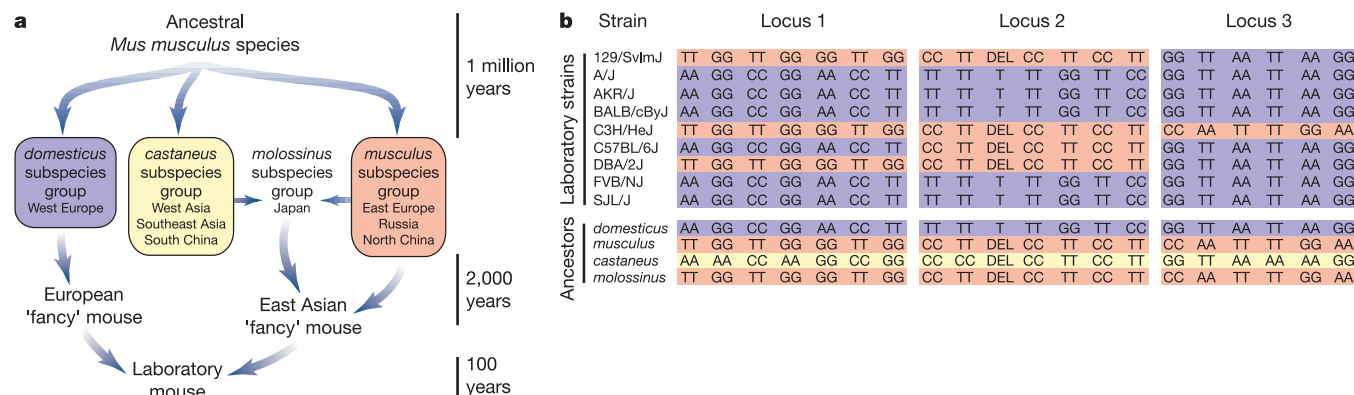


Figure 1 History of the inbred laboratory mouse and the resulting patterns of genetic variation. **a**, Genetic contributions are expected from both European (*domesticus*) and Asian (*musculus* and *castaneus*) subspecies because the mice were derived from Asian or European 'fancy' stocks collected by a few mouse suppliers. **b**, Most loci show two haplotypes and match wild-derived representatives of *domesticus* (WSB/Ei) (purple),

musculus (CZECH/Ei) (orange) or *molossinus* (MOLF/Ei) but not *castaneus* (CAST/Ei) (yellow). Chromosome 14:121 Mb (locus 1), chromosome 6:139 Mb (locus 2) and chromosome 2:107 Mb (locus 3) are represented by polymorphisms from 1-kb windows re-sequenced in nine strains.

sequence from strain 129 (the strain with the most finished sequence available after B6). A total of 59 finished sequence contigs from strain 129 (various substrains and clone lines including 129, 129/Sv, 129SvEvTac, 129Ola, RPCI_21, RPCI_22, MGS1 and CITB_CJ7) ranging in size from 27 kb to 1.7 Mb and totalling 17 Mb, were obtained from GenBank and compared with the B6 genome assembly as described above. Considering only the first 50 kb of each sequence (to normalize the sequence contributions of different genomic regions in the comparison), a distinct bimodality of polymorphism rate is revealed (Fig. 2). The observed distribution of SNP rates in 50-kb segments is strikingly nonrandom when compared with the expected distribution, assuming a constant polymorphism rate (14 SNPs per 10 kb, the overall observed rate), consistent with the observations of nonrandomness made in a previous small-scale study⁹. Of the regions examined, 39% ($\pm 6.3\%$) fall into the higher mode of this distribution (more than 10 SNPs per 10 kb) and have a mean of 36 SNPs per 10 kb. Those in the lower mode (less than 5 SNPs per 10 kb) have an average polymorphism rate of 1.6 SNPs per 10 kb, suggesting that these 58% ($\pm 6.4\%$) of genomic segments share a much more recent co-ancestry between B6 and 129. Because of existing evidence of considerable diversity caused by contamination of various substrains of 129 (refs 10–12) we separated individual substrains and clone libraries (CITB-CJ7-129Sv, RPCI21-129S6/SvEvTac, RPCI22-129S6/SvEvTac and MGS1-129/Ola) but see the same general bimodality in each when compared with B6.

As in the B6–B6 comparison, we selected and re-sequenced 15 putative 129–B6 SNPs from distinct genomic regions with very low SNP rates. Only four of these 15 were determined to be true polymorphisms. This suggests that there are a few true polymorphisms in these low-SNP-rate regions, but that the actual rate is probably closer to one per 20,000 bp (~ 0.5 SNPs per 10 kb). By contrast, 79 of 80 candidate B6–129 SNPs selected from regions with a high SNP rate were found to be true SNPs by validation through

re-sequencing. Because nearly all SNPs ($>95\%$) detected in this study come from regions of high SNP rate, this data indicates an overall high validation rate for SNPs discovered in this survey.

Low-SNP-rate regions in this two-strain comparison often appear to persist over considerable distances in the finished sequence (most often the full length of the sequence). By examining those finished sequences that have a low or absent SNP rate in the first 50 kb, we find only three of 32 (9%) with a high SNP rate (>10 SNPs per 10 kb) in the second 50 kb of the sequence and four of 28 (14%) with a high SNP rate in the third 50-kb segment (both $P < 0.01$ when compared with the overall average of 37% of 50-kb segments that have a high SNP rate). Although the number of finished sequences longer than 150 kb was limited, these data suggest that segments of high or low polymorphism rate between strains extend over long genomic distances. In a few cases, long finished sequences contain an overall intermediate SNP rate, which reflects a sharp transition between a low and high SNP rate region (Fig. 3); this transition emphasizes the bimodal nature of polymorphism rate in these genome comparisons. Only a handful of finished sequences showed such transitions, precluding genome-wide generalizations, but qualitative observation of the transitions suggests that they are sharp, rather than diffuse, in character.

Having discovered a strong bimodality distinguishing regions of low and high polymorphism rate over long genomic segments, we then extended the observations to the entire genome and to several strains. We used whole genome shotgun (WGS) reads for three strains of inbred mice to examine patterns of variation on a genome-wide scale; the strains were: 129S1/SvImJ (Jackson Laboratories stock 002488, henceforth 129) with 119,232 reads, C3H/HeJ (stock 000659, henceforth C3H) with 68,160 reads, and BALB/cByJ (stock 001026, henceforth BALB) with 38,400 reads. SNP detection was carried out once more with SSAHA-SNP⁸, using the WGS reads that could be uniquely placed on the B6 assembly³. A total of 79,269 candidate SNPs were found across all strains. The observed distribution in each strain follows the general pattern seen when using the finished sequences, with many long regions having either no SNPs (for example, a 26-Mb region on chromosome X contained 131 consecutively placed reads with no SNPs) or many SNPs (for example, a 5-Mb segment of chromosome 12 contained 36 consecutively placed reads with one or more candidate SNPs).

We would not expect regions of low-SNP rate to contain

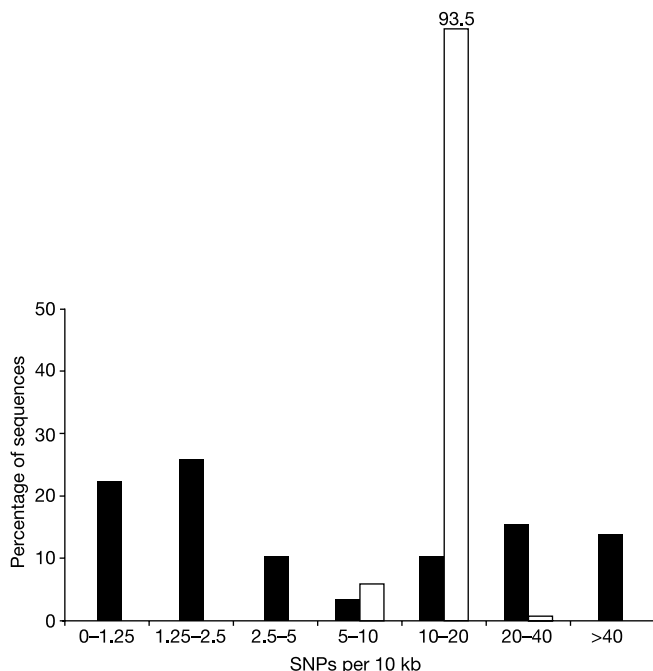


Figure 2 SNP frequency in the first 50 kb of finished sequences for strains 129 as compared to the B6 assembly. We note the bimodal distribution of observed SNPs (black) and the difference from the expected Poisson distribution (white) with an average of 14 SNPs per 10 kb (the observed rate across the entire data set). This indicates a very non-uniform distribution of SNPs in the genome with the presence of regions with a high SNP rate (~ 45 SNPs per 10 kb) and regions with a low SNP rate (~ 0.5 SNPs per 10 kb).

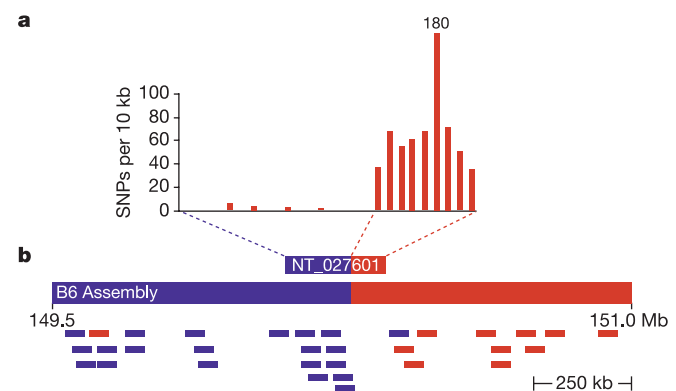


Figure 3 A clear transition from low to high SNP rate between strains 129 and B6 in a 1-Mb region at the telomeric end of chromosome 5. **a**, Finished sequence NT_027601 when aligned against the B6 whole genome assembly demonstrates a low-SNP-rate region (the first 161 kb) followed by a high-SNP-rate region (the last 101 kb) and an abrupt transition between SNP rate regions. **b**, Whole genome shotgun reads from the same genomic region provide evidence of the same well-defined transition point. Bars correspond to 500-bp reads containing one or more SNPs (red), or no SNPs (blue) when compared with the B6 assembly.

exclusively reads with no candidate SNPs, if only because of the inherent sequencing error rate. Similarly, regions of high SNP rate will certainly contain individual reads with no SNPs because the average rate in such regions is only about two SNPs per 500-bp read. We therefore developed a hidden Markov model (HMM)¹³ to use the WGS read data to parse the genome. The approach assumes that there are two types of genomic regions (of low SNP rate and high SNP rate, as suggested by the finished sequence analysis) and uses all reads on a chromosome to identify those genomic segments that have an unambiguously high or low SNP rate. In the 129–B6 comparison, more than 90% of the genome is contained in such unambiguous segments (high or low) and 50% of the genome is contained in defined segments of greater than 2.0 Mb in length (the N50 value). Although shotgun sampling may occasionally miss short segments and will not capture the discrete nature of transitions (as in Fig. 3), the available 70,000 placed reads allow an accurate estimate of the N50 segment size (Supplementary Fig. 1). With a random distribution of transitions, an N50 value of 2.0 Mb suggests an average block size of 1.2 Mb. SNP-poor ('low-SNP') regions account for 67% of the genome in the comparison between strains 129 and B6. The overall SNP rate found in high-SNP regions is 45 per 10 kb, and the rate in low-SNP regions is 1.0 per 10 kb. The fraction of genome in low- and high-SNP-rate regions and the SNP rate within these regions are very similar for the C3H and BALB comparisons to B6 and attests to the common origins of all four strains (Supplementary Fig. 1).

Such long regions of either recent or distant co-ancestry should, once identified, be clearly distinguishable from one another using data from existing SSLP maps. Regions identified as 'high' or 'low' in the HMM analysis were cross-referenced with SSLP data from the MIT genetic map¹⁴ (using 2,605 SSLPs that were uniquely mapped onto the genome assembly³ and for which SSLP allele size information was available for B6, C3H and BALB). In pairwise comparisons between B6 and either C3H or BALB, SSLP polymorphism rates correlate strongly with the low- and high-SNP-rate regions assigned by the HMM: 31% (low) versus 75% (high). The fact that SSLPs are polymorphic in SNP-poor regions attests to their high mutability even in regions of recent co-ancestry and to the difficulty of recovering ancestral patterns directly from SSLP data.

To examine haplotype variation across several inbred laboratory strains, we selected 27 genomic segments (500–1,000 bp in length); nine from each of the three strains (C3H, 129 and BALB) in which the shotgun SNP discovery had revealed 4–15 candidate SNPs between the sequenced strain and B6. These segments were then completely sequenced in a total of nine inbred strains: B6, 129, C3H, BALB, A/J, AKR/J, DBA/2J, FVB/NJ and SJL/J (examples in Fig. 1b). In 22 of the 27 regions, only two different haplotypes were observed—emphasizing that the founder population from which these strains have been derived was extremely limited. The ancestry of these haplotypes was classified by sequencing representative wild-

derived inbred lines of *M. m. domesticus* (WSB/Ei), *M. m. musculus* (CZECHII/Ei), *M. m. molossinus* (MOLF/Ei), and *M. m. castaneus* (CAST/Ei). Even in this extremely limited sampling of the ancestral populations, 67% of the haplotypes can be readily ascribed to the *domesticus* line and another 21% appears to derive from the Asian mice (primarily *musculus* and *molossinus*, which are usually identical). Despite comparing the inbred strains to only a single isolate from each heterogeneous ancestral population, this straightforward experiment appears to support the historical expectation (Fig. 1) completely. To examine the extent of these ancestral haplotypes, we sequenced these same strains at eight pairs of 1-kb windows separated by 100–200 kb that occurred in long high-SNP-rate segments identified in the original B6–129 finished sequence comparison. In 49 of 56 cases, identity in the seven other inbred strains (excluding B6 and 129) to either B6 or 129 persisted over the entire segment, providing an estimate (through linkage disequilibrium) consistent with 1-Mb blocks. In the other cases, the identity to B6 or 129 did not persist between segments, suggesting an historical recombination had taken place during the creation of the inbred strains.

The present observations of juxtaposed regions of consistently high and low diversity reveal the impact of breeding history on diversity in the mouse genome. In comparisons of any inbred laboratory strain with B6, we find two-thirds of the genome to have a very low polymorphism rate, indicating genomic regions in which the two strains share a very recent common subspecies origin (usually both *domesticus* or both *musculus*) (Fig. 4). By contrast, the remaining one-third of the genome in these comparisons shows considerable divergence (1 SNP every 200 bp), indicating distant ancestry consistent with a previous measure of genome-wide diversity between the laboratory strains and *M. m. castaneus*⁹ (*M. m. musculus* and *M. m. castaneus* are roughly equidistant from *M. m. domesticus*^{2,15}), suggesting that these represent regions in which the two strains inherited the region from different subspecies (most often, one *domesticus* and the other *musculus*). Finally, a more detailed look at individual regions in these inbred laboratory strains suggests that these patterns represent the effects of recent breeding practice that has created an admixture of two major ancestral sources (Fig. 1).

The segmented nature of these genomes and the lack of heterogeneity within the segments provides important insights into the interpretation of quantitative trait mapping experiments using inbred strains and offers the potential for the acceleration of

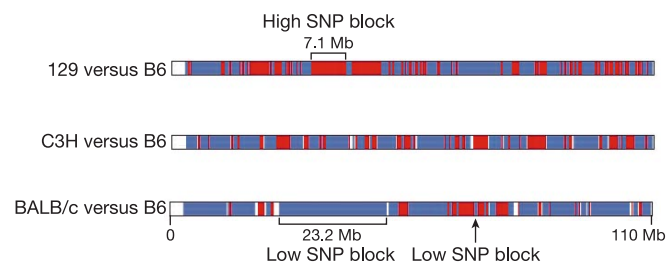


Figure 4 Mosaic structure of the mouse genome. Chromosome 15 comparisons between 129S1/SvImJ, C3H/HeJ and BALB/cByJ with C57BL/6J reveal regions of low SNP rate (~0.5 SNPs per 10 kb, blue) and regions of high SNP rate (~45 SNPs per 10 kb, red). Half of the genome is contained in segments (low or high) of greater than 2 Mb in length. Low-SNP-rate regions define roughly two-thirds of each genome comparison.

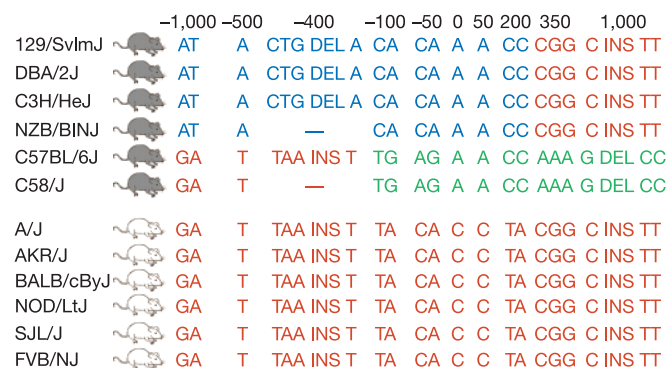


Figure 5 Association between a single haplotype and the albinism phenotype caused by a mutation at the tyrosinase locus²⁰. Columns show SNPs discovered in ten 500-bp assays with positions (kb) relative to the centre of the genomic segment containing the gene (GenBank accession GI:12852585). The causal mutation (Cys103Ser) is located at +32.6 kb. The association between phenotype and ancestral haplotype for 12 strains would be sufficient to identify a haplotype background and 'critical region' of ~500 kb (including the assays from -100 kb before Tyr to 200 kb after Tyr) likely to contain the albinism mutation.

subsequent positional cloning. For many phenotypes, positional cloning after initial quantitative trait locus (QTL) mapping should now focus on subregions where the parental strains are recognized to differ, as more than 95% of the total genetic variation will be found in that one-third of the genome. This can be much more powerful in cases where a locus has been mapped in several strain combinations. Several parental strains could then be simultaneously compared to discover an even smaller critical region. In addition, the observation of the same QTL in several strain combinations makes it even more likely that the QTL is due to ancestral differences common to many strains rather than a rare mutation specific to a single strain. For phenotypes not subject to strong selection (for example, otherwise phenotypically benign modifiers of knockouts, transgenes and spontaneously arising mutations), we should be able to use the parental phenotype data of many inbred strains to correlate ancestral patterns directly with phenotype.

As an example, we constructed haplotypes from 12 inbred strains across 2 Mb surrounding the *tyrosinase* gene on chromosome 7. Figure 5 shows the clear correlation of ancestral haplotype and the phenotype (albinism) due to a mutation in this gene. Without knowing the gene or functional polymorphism, the pattern of ancestral haplotypes in only 12 strains would suffice to identify the short genomic segment (~500 kb) surrounding the gene as the region likely to contain the albinism locus had we initially identified the chromosome 7 region in a traditional mapping cross. A much larger panel of strains could reduce the critical region further and would provide much stronger statistical evidence for association of genomic segment to phenotype. In effect, existing laboratory strains serve as a natural set of recombinant inbred lines—although with more than ten times as many breaks as in conventional recombinant inbred strains. A detailed understanding of the ancestral origins of haplotypes should allow researchers to use the entire panel of inbred laboratory strains in this fashion.

A genome-wide characterization of the segmented genetic variation of the 50–100 commonly used inbred mouse strains would thus enrich the utility of large-scale phenotyping efforts currently underway¹⁶ (such as the Mouse Phenome Project¹⁷, <http://www.jax.org/phenome>). As shown, it can in some cases allow *de novo* mapping of major loci to small candidate regions simply on the basis of correlation between ancestral segment identity and phenotype. In addition, it will significantly improve our ability to recognize appropriate strain combinations for confirmatory mapping of QTLs (those where genetic ancestry as well as phenotype differs) and for mapping modifiers (those with the greatest phenotypic differences among strains known to share ancestral identity at major loci). A haplotype map offering a precise picture of the history of each individual genomic region is likely to become a valuable complement to traditional approaches in genetic mapping. □

Methods

Finished sequences

Single or merged finished bacterial or phage artificial chromosome sequences for *Mus musculus* were retrieved from the National Center for Biotechnology Information web site (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/NTStrainList). Those of either solely strain 129 and its derivatives, or solely strain B6 were extracted from the Entrez database at NCBI.

SNP discovery

The MGSCv3 C57BL/6J whole genome shotgun assembly, generated from over 40 million paired-end reads and assembled using Arachne¹⁸, was compared with finished sequences (incorporating quality scores) in 1,000-bp windows using SSAHA-SNP⁸. Uniquely placed sequences with a minimum base quality score of Phred 23 (ref. 19) and a maximum of twenty SNPs per 1,000-bp window were included.

SNP rate

The SNP rate for each 50-kb segment of finished sequence was assessed by taking the total of SNPs divided by the total of high-quality bases examined. Individual finished sequences were grouped by SNP rate (0–1.25, 1.25–2.5, 2.5–5, 5–10, 10–20, 20–40, >40 SNPs per 10 kb) within each strain and compared to a Poisson expectation based on the overall observed SNP rate.

SNP validation

Fifteen putative SNPs from the C57BL/6J assembly versus finished C57BL/6J sequence discovery were re-sequenced in 17 DNAs. The DNAs represented males and females from seven different C57BL/6J founderlines kept at the Jackson Laboratories and the C57BL/6J DNA remaining at the Whitehead Institute after construction of the WGS libraries. The SNP loci were amplified and sequenced using fluorescent M13 dye-primer sequencing on an ABI 377 Sequencer (Applied Biosystems; for details see Supplementary Information).

Validation as above was carried out for putative SNPs from low-rate (15 SNPs) and high-rate (80 SNPs) regions detected in the B6–129 finished sequence comparison.

Whole-genome shotgun read analysis

119,232, 68,160 and 38,400 random shotgun reads (paired-ends from 4-kb plasmids) for strains 129S1/SvImJ, C3H/HeJ and BALB/cByJ respectively were found as part of the Mouse Genome Sequencing Consortium³ SNP discovery. Vector- and quality-trimmed reads were assessed for 20-base windows having an average quality score below Phred 20, implying an error rate of lower than 1 base miscalled per 100 bases¹⁹. The longest block of sequence between any two windows was passed onto SNP discovery if that block exceeded 250 bp. SNP detection was by SSAHA-SNP⁸ using 500-bp windows and a maximum of ten SNPs per window. The number of reads passing our quality thresholds, SSAHA-SNP, and post-SSAHA filtering for repetitive sequence were 67,974, 34,949 and 19,686 for strains 129S1/SvImJ, C3H/HeJ and BALB/cByJ respectively.

Determination of SNP low- and high-rate regions

A hidden Markov model (HMM) fitted a two-state model of low polymorphism rate (low) and high polymorphism rate (high) using the forward-backward algorithm¹³. Prior estimates for the observational and transitional probabilities used in the HMM were obtained by computational observation of frames of three consecutive reads on either side of the read to be evaluated.

Regions of high and low polymorphism were defined as continuous segments estimated to have greater than 0.95 or less than 0.05 probability of being 'high' according to the HMM, respectively. Block sizes were the distances between the first base for the first read and the last base for the final read where consecutive reads met 'low' or 'high' criteria.

Re-sequencing of multiple strains for haplotype studies

To determine the number and ancestry of haplotypes among inbred strains a panel of nine inbred strains (C57BL/6J, 129S1/SvImJ, C3H/HeJ, BALB/cByJ, A/J, AKR/J, DBA/2J, FVB/NJ and SJL/J) and four wild-derived inbred ancestral strains (*M. m. domesticus* WSB/Ei, *M. m. musculus* CZECHII/Ei, *M. m. molossinus* MOLF/Ei and *M. m. castaneus* CAST/Ei) were re-sequenced as above (see 'SNP validation' section). A total of 27 random 1-kb regions (nine from the 129 strain, nine from C3H and nine from BALB/c discovery) containing 4–11 SNPs were re-sequenced and searched for SNPs. Haplotypes were inferred by inspection at each locus. Haplotypes varying by only one SNP between inbred strains were counted as the same haplotype. The number of haplotypes at each locus was tallied and compared with the haplotypes in the ancestral mice.

Eight paired 1-kb regions derived from the finished 129 sequence comparison were re-sequenced in the same strains. The paired regions all resided 100–200 kb apart in high-SNP regions. Haplotypes were scored and counted as above. The observation of 87% of paired loci separated by an average of 150 kb with concordant haplotype suggests (by Poisson) an average block size of 1.1 Mb.

Raw data from SNP discovery and re-sequencing is available at <http://www-genome.wi.mit.edu/~mjdaly>.

Received 12 September; accepted 28 October 2002; doi:10.1038/nature01252.

1. Silver, L. M. *Mouse Genetics* (Oxford Univ. Press, New York, 1995).
2. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nature Genet.* **24**, 23–25 (2000).
3. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* (this issue).
4. Bishop, C. E., Boursot, P., Baron, B., Bonhomme, F. & Hatat, D. Most classical *Mus musculus domesticus* laboratory mouse strains carry a *Mus musculus musculus* Y chromosome. *Nature* **325**, 70–72 (1985).
5. Yonekawa, H. *et al.* Relationship between laboratory mice and subspecies *Mus musculus domesticus* based on restriction endonuclease cleavage patterns of mitochondrial DNA. *Jpn. J. Genet.* **55**, 289–296 (1980).
6. Ferris, S. D., Sage, R. D. & Wilson, A. C. Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. *Nature* **295**, 163–165 (1982).
7. Bonhomme, F., Guénet, J.-L., Dod, B., Moriawaki, K. & Bulfield, G. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *J. Linn. Soc.* **30**, 51–58 (1987).
8. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
9. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
10. Simpson, E. M. *et al.* Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nature Genet.* **16**, 19–27 (1997).
11. Schalkwyk, L. C. *et al.* Panel of microsatellite markers for whole-genome scans and radiation hybrid mapping and a mouse family tree. *Genome Res.* **9**, 878–887 (1999).
12. Threadgill, D. W., Yee, D., Matin, A., Nadeau, J. H. & Magnuson, T. Genealogy of the 129 inbred strains: 129/SvJ is a contaminated inbred strain. *Mamm. Genome* **8**, 390–393 (1997).
13. Rabiner, L. A. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
14. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
15. Prager, E. M., Orrego, C. & Sage, R. D. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150**, 835–861 (1998).

16. Threadgill, D. W., Hunter, K. R. & Williams, R. W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* **13**, 175–178 (2002).
17. Paigen, K. & Eppig, J. T. A mouse phenotype project. *Mamm. Genome* **11**, 715–717 (2000).
18. Batzoglou, S. *et al.* ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12**, 177–189 (2002).
19. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. 2. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
20. Yokoyama, T. *et al.* Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. *Nucleic Acids Res.* **18**, 7293–7298 (1990).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank the Mouse Genome Sequencing Consortium for sequencing and SNP discovery efforts and the University of Queensland for providing study leave to C.M.W. We thank W. Frankel and K. Scott for providing DNA from seven B6 founder stocks from the Jackson Laboratories. We also thank R. Jaenisch, D. Page, A. Chess, W. Dietrich, J. Hirschhorn, D. Altschuler, J. Barrett and M.-P. Reeve for comments on the manuscript; B. Gilman, S. Schaffner and E. Karlsson for computational assistance; and L. Gaffney for assistance with figures. M.J.D. is supported through a computational biology fellowship funded by Pfizer, Inc.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.D. (e-mail: mjdaly@genome.wi.mit.edu) or K.L.T. (e-mail: kersli@genome.wi.mit.edu).

Numerous potentially functional but non-genic conserved sequences on human chromosome 21

Emmanouil T. Dermitzakis*, **Alexandre Reymond***, **Robert Lyle***, **Nathalie Scamuffa***, **Catherine Ucla***, **Samuel Deutsch***, **Brian J. Stevenson†‡**, **Volker Flegel†‡**, **Philipp Bucher†§**, **C. Victor Jongeneel†‡** & **Stylianos E. Antonarakis***

* Division of Medical Genetics, 1 Rue Michel-Servet, University of Geneva Medical School and University Hospitals of Geneva, CH-1211 Geneva, Switzerland
† Swiss Institute of Bioinformatics, § Swiss Institute for Experimental Cancer Research, and ‡ Office of Information Technology, Ludwig Institute for Cancer Research, 155 Chemin des Boveresses, CH-1066 Epalinges, Switzerland

The use of comparative genomics to infer genome function relies on the understanding of how different components of the genome change over evolutionary time^{1–3}. The aim of such comparative analysis is to identify conserved, functionally transcribed sequences such as protein-coding genes and non-coding RNA genes, and other functional sequences such as regulatory regions^{4,5}, as well as other genomic features. Here, we have compared the entire human chromosome 21 with syntenic regions of the mouse genome, and have identified a large number of conserved blocks of unknown function. Although previous studies have made similar observations^{6,7}, it is unknown whether these conserved sequences are genes or not. Here we present an extensive experimental and computational analysis of human chromosome 21 in an effort to assign function to sequences conserved between human chromosome 21 (ref. 8) and the syntenic mouse regions. Our data support the presence of a large number of potentially functional non-genic sequences, probably regulatory and structural. The integration of the properties of the conserved components of human chromosome 21 to the rapidly accumulating functional data for this chromosome^{9,10} will improve considerably our understanding of the role of sequence conservation in mammalian genomes.

The sequence of human chromosome 21 (ref. 8) was obtained from the National Center for Biotechnology Information (NCBI) and aligned with PipMaker¹¹ to the mouse orthologous sequences

(both sequences were hard-masked with Repeatmasker). Details and parameters of the alignment are described in the Methods. Briefly, 33.5 megabases (Mb) of human chromosome 21 sequence was compared to approximately 21 Mb of mouse sequence from chromosomes 16, 17 and 10 (refs 7, 12). A large number of ungapped conserved sequences were identified and their

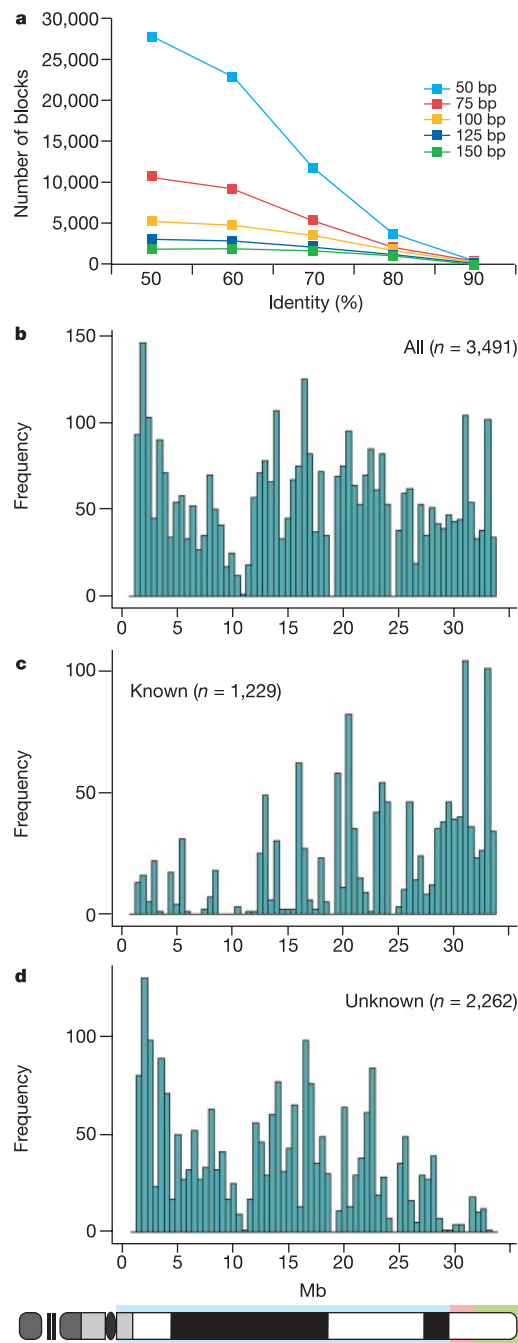


Figure 1 Distribution of conserved blocks on the 33.5 Mb of human chromosome 21 (long arm). **a**, Number of conserved sequence blocks identified (known and unknown) as a function of the threshold criteria for the size and percentage identity of the ungapped conserved blocks. **b–d**, The distribution of conserved blocks on the 33.5 Mb of human chromosome 21 at ≥ 100 bp and $\geq 70\%$ identity is shown. Histograms showing all 3,491 conserved blocks (**b**), 1,229 conserved blocks corresponding to known exonic sequences (**c**), and 2,262 conserved blocks of unknown function (**d**). Below the histograms, human chromosome 21 with its banding pattern, and with the corresponding mouse syntenic regions is shown (mouse chromosomes 16, 17 and 10 are indicated in blue, pink and green, respectively).

distribution depending on the percentage identity and size thresholds is shown in Fig. 1a. We chose to analyse ungapped sequences of ≥ 100 base pairs (bp) with $\geq 70\%$ identity, yielding a total of 3,491 blocks. These threshold criteria are informative for the identification of important genomic functional elements^{13,14}. From those 3,491 blocks, 1,229 correspond partly or fully to known exonic sequences and annotated pseudogenes of human chromosome 21 (denoted as known)^{8,10,15–18}, and the remaining 2,262 blocks have unknown function (denoted as unknown). When compared with each other they did not have significant similarity, indicating that they are not repeats or widespread structural features. They also appear to be single copy in the human genome. Figure 1b–d shows the distribution of conserved blocks along the long arm of human chromosome 21. A large number of the conserved sequences are present in the gene-poor region of human chromosome 21 (ref. 8), suggesting that this region is not a ‘functional desert’.

Unknown conserved blocks probably belong to the following six categories with respect to their function: (1) alternatively spliced, unknown exons of known genes; (2) exons of unknown genes; (3) non-coding RNAs; (4) *cis*-regulatory regions; (5) functional sequences of unknown significance; and (6) non-functional sequences with low divergence due to low substitution rate. In this study we investigated how the 2,262 unknown conserved blocks are distributed in the above six categories of conserved sequences.

To test experimentally the coding potential of the unknown 2,262 conserved blocks we applied four different methods to obtain candidate gene models in the human sequence: GrailEXP, Pro-Gen¹⁹, human and mouse expressed sequence tag (EST) matches and adjacent conserved blocks. We also used NCBI annotations mostly on the basis of the gene prediction program GenomeScan, and performed BLAST analysis to identify which of the conserved blocks matched NCBI annotations. Out of 454 hypothetical loci in the NCBI human chromosome 21 annotations, most of which are GenomeScan predictions, only 18 matched 59 unknown conserved blocks. Furthermore, a total of 123 gene models (24 GrailEXP predictions, 10 Pro-Gen predictions, 26 EST matches and 63

adjacent conserved blocks) were tested experimentally. Oligonucleotides spanning putative introns of the gene models were used for polymerase chain reaction with reverse transcription (RT–PCR) amplification and sequencing on a panel of complementary DNA pools from 20 human adult and fetal tissues. Out of 123 models, only 2 spliced transcripts (1 Pro-Gen model and 1 GrailEXP model) were confirmed (1.7%). By using this RT–PCR methodology in other studies we documented transcripts of more than 98% of the human chromosome 21 genes, and discovered 19 additional transcripts^{10,18}; we also found 139 additional transcripts in the whole mouse genome¹². All 26 EST and 63 adjacent block models were also tested in the presence and absence of reverse transcriptase in hepatoma cell line and brain cDNA pools, to reveal possible unspliced non-coding RNAs. None of the 89 models showed evidence for the presence of an unspliced transcript in the tissues tested. In addition, using QRNA²⁰, which predicts coding sequences and non-coding RNAs on the basis of the pattern of substitution between two sequences, only 225 and 193 conserved blocks were predicted as coding or non-coding RNAs, respectively.

To investigate further the potential of functional transcription of the conserved blocks, we analysed a data set that reported the levels of transcriptional activity across the human chromosome 21 unique genomic sequence in 11 human cell lines, using Affymetrix oligonucleotide arrays²¹. We first created a non-redundant set of the oligonucleotide sequences reported as positive in at least one cell line²¹, and then used BLAST to identify those sequences that matched conserved blocks. To avoid spurious positives we selected conserved blocks that corresponded to at least two positive oligonucleotides of the array. A total of 485 (21.4%) conserved blocks matched two or more positive oligonucleotides in the array.

We combined 6 pieces of computational and experimental evidence (GrailEXP prediction, GenomeScan prediction, human EST match, mouse EST match, QRNA prediction-coding, and RNA- and oligonucleotide array match) to obtain a signal of functional transcription for each of the 2,262 unknown blocks. Figure 2 shows that this signal is weak and that most of the blocks (63%) have no

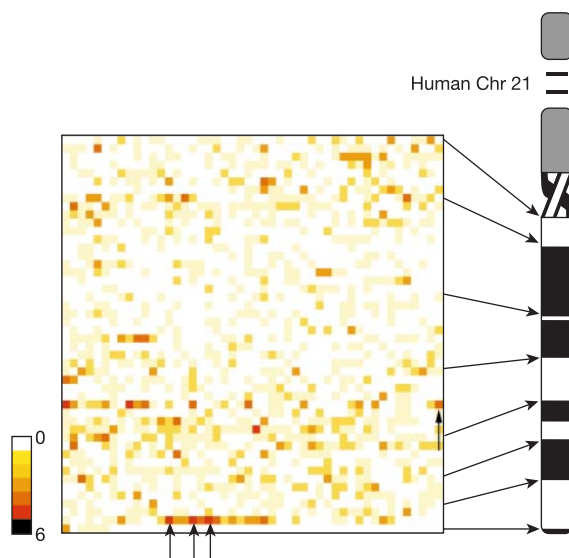


Figure 2 Colour-coded panel of the evidence of functional transcription (scale: 0 to 6) in 2,262 unknown conserved blocks. Arrows indicate the position of the last block of each row on human chromosome 21 (far right). The blocks are depicted as squares and the order along human chromosome 21 (centromere to telomere) is left to right and up to down. The 6 criteria are: GrailEXP, GenomeScan, human EST, mouse EST, QRNA (coding or RNA) and microarray. Vertical arrows indicate NCBI annotations with high support (bottom) and a gene identified during this analysis (right). Note the contrast of low support for most blocks compared with the blocks indicated by arrows.

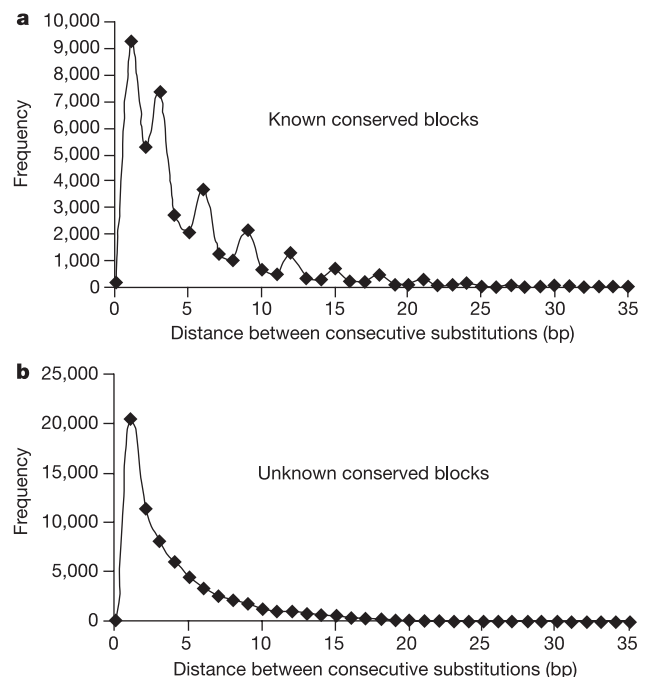


Figure 3 Frequency distribution of the distances between consecutive nucleotide substitutions in known (a) and unknown (b) conserved blocks. Note the periodicity (period = 3) in a due to the high frequency of synonymous substitutions at the third codon position, which is absent from b.

support. Three conserved blocks that probably correspond to genes with high experimental support—according to the NCBI annotation—and a gene identified by our group during this analysis stand out as highly supported in Fig. 2 (vertical arrows). (Supplementary Table 1 summarizes the coding potential and expression characteristics of all 2,262 unknown conserved blocks as derived by the *in silico* and experimental analysis.) The pattern seen in Fig. 2 combined with the previously described experimental RT-PCR analysis suggests that the functional transcription potential of

most of the 2,262 unknown blocks is low.

None of the above methods (except for QRNA) make use of the pattern of substitution between the human and mouse sequences. In coding sequences most of the nucleotide changes are synonymous and occur at the third position of a codon. To exploit that characteristic²², for each of the human–mouse alignments in the 2,262 unknown blocks and the 1,229 known (exonic) blocks we obtained measures of the distances (in bp) between consecutive nucleotide substitutions. Figure 3 shows the distribution of these distances in known and unknown conserved blocks. Within the known blocks there is a pronounced periodicity (period = 3) owing to the high frequency of synonymous changes in the third position, which is not at all present in the unknown set. We performed the same analysis in the unknown blocks identified by GrailEXP, GenomeScan, oligonucleotide array data, and QRNA, and none showed periodicity (see Supplementary Information). These data further demonstrate that most of the unknown conserved blocks are unlikely to be coding sequences.

To test for active (functional) conservation we attempted to amplify by PCR a total of 220 (about 10%) of the unknown conserved blocks in a third species, the rabbit. This species was chosen because, on the basis of mammalian phylogeny, it is almost equally distant to mouse and human (slightly closer to mouse). Thus there has been enough time since its separation from the two other species to allow efficient phylogenetic footprinting^{23–25}. The distribution of conserved blocks amplified in rabbit is very similar to the distribution of the unknown conserved blocks on human chromosome 21. Out of the 220 unknown conserved blocks, we obtained 111 distinct rabbit sequences without any bias in their position on the chromosome (remaining sequences did not amplify or had multiple PCR products), and a total of 18,288 nucleotides were aligned in all three species with MultiPipMaker. Average measures of divergence of the rabbit sequences were within the expected level (Supplementary Information), which supports the hypothesis that these sequences are orthologous. Three-way species sequence analysis²⁶ allows for a detailed description of the pattern of substitution within these regions. Although in ref. 6 some three-way analysis was performed, the methodology did not allow for either reliable quantification of conservation or analysis of the pattern of substitution, which can reveal significant properties of these sequences.

To perform a fine analysis of substitution patterns, we counted the species-specific substitutions in each of the 111 conserved blocks (see Supplementary Information for methods) and constructed a 3×111 contingency table to test whether there is heterogeneity in the substitution pattern. A Fisher's exact test shows that the pattern is significantly heterogeneous ($P < 10^{-4}$). All three species seem to contribute to this heterogeneity (Fig. 4a) and thus it is not an artefact of the presence of paralogues in rabbit data or differentially accelerated substitution rate in mouse. Performing the test in a sliding window of blocks within 1 Mb or 500 kb of genomic sequence, most tests show $P < 0.001$; this heterogeneous pattern cannot be explained by regional variation of species-specific mutation or substitution rate. Therefore, despite the fact that these sequences are conserved as a whole between species, the heterogeneous species-specific substitution patterns suggest differential selective pressure. Fine scale species-specific changes in a conserved sequence could indicate differences in putative functional regions (for example, regulatory) that contribute to species differences^{5,27}.

A recent study suggested that variation in substitution rate among genomic regions²⁸ is probably due to different nucleotide composition bias between species in some genomic regions²⁹, which would increase the mutational bias of a sequence within these regions. This allows for a prediction of the effect of the genomic context to the neutral substitution rate. We explored the substitution pattern of the conserved regions by focusing on the type of nucleotide changes that have occurred. We calculated the difference in the (G+C) content between human and mouse in the

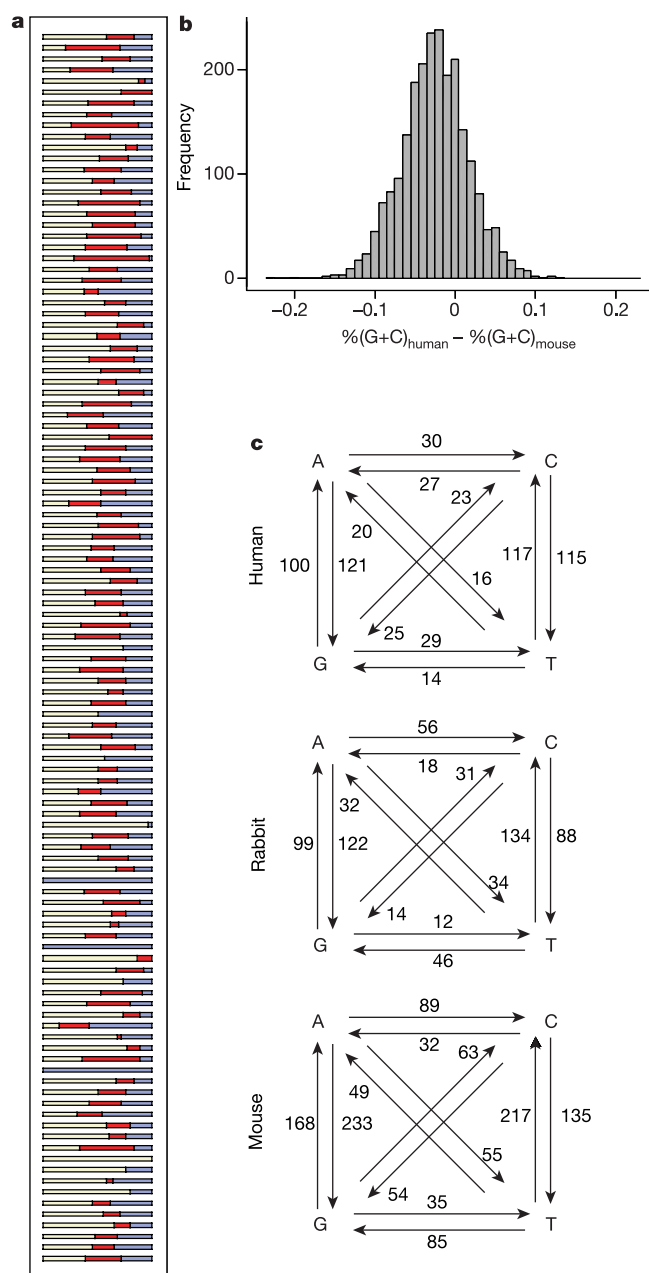


Figure 4 Analysis of sequence comparison of three species. **a**, Proportional representation of the species-specific nucleotide substitutions in human (blue), rabbit (red) and mouse (yellow). Each bar represents one of the 111 regions sequenced in rabbit, ordered by their position on human chromosome 21 (centromere to telomere). Note the apparent heterogeneity. **b**, Distribution of the difference in the percentage (G+C) content between human and mouse ($\%(G+C)_{\text{human}} - \%(G+C)_{\text{mouse}}$). The trend towards higher (G+C) content in mouse is obvious by the shift of the distribution to the left. **c**, Species-specific substitution counts in human, mouse and rabbit. The (A+T) to (G+C) bias in mouse and rabbit is evident.

2,262 conserved blocks (Fig. 4b). There is a pronounced skew towards higher (G+C) content in the mouse sequence, which implies different mutational pressure in this genomic region in the two species. With the three-way species alignment we can determine the direction of these substitutions (Fig. 4c) and decide which of the species drive the (G+C) content differential in the conserved regions. Out of the 637 human-specific substitutions 293 (46%) were A+T to G+C and 271 (42.5%) were G+C to A+T. For mouse these numbers were 624 A+T to G+C out of 1,215 (51.3%) and 370 G+C to A+T out of 1,215 (30.4%); for rabbit these were 358 A+T to G+C out of 686 (52.1%) and 217 G+C to A+T out of 686 (31.6%). This demonstrates that there is no mutational bias on human chromosome 21, whereas there is an A+T to G+C mutational bias in the mouse and rabbit syntenic regions. The difference in the mutational bias between human and mouse/rabbit in these regions increases the pressure for substitutions so as to reach the species-specific nucleotide composition equilibrium. This observation suggests the presence of an extra force of nucleotide change (A+T to G+C bias) that increases the substitution rate relative to a non-bias model. High sequence conservation with the presence of this bias is an additional support for their functional significance.

A preliminary description of sequence conservation along human chromosome 21 has been presented in previous studies^{6,7}. These studies have detected patterns of conservation similar to our study, but the question of whether these unknown conserved sequences are unannotated genes or are non-genic remained unanswered. This is an important question because genic and non-genic sequences require different experimental approaches for their functional analysis. The fact that many conserved sequences are not present in the current gene annotation does not necessarily mean that they are non-coding, as discussed in the accompanying manuscript, where a significant number of new genes have been verified experimentally¹². In our study, we tested specific hypotheses about the importance of the unknown conserved sequences. With this analysis we rejected the hypothesis that the unknown conserved sequences on human chromosome 21 are unannotated genes, therefore emphasizing the need for specific experimental strategies to unravel the function of such non-coding sequences.

Another major aspect of the present study is the use of direct nucleotide alignment to perform the analysis. One study used high-density microarrays⁶, which provide qualitative but not reliable quantitative data in terms of numbers and types of substitutions at levels of divergence such as those between human and mouse. Another study performed nucleotide comparisons⁷, but the data were restricted to the degree of overlap of the conserved sequences with the known genes, and only involved two species. In our study we describe certain characteristics of the pattern of substitutions between human, mouse and rabbit that allowed us to test the coding potential of the conserved sequences, as well as the degree of active conservation. Our analysis demonstrates the power of combining computational, experimental and evolutionary analysis for the understanding of genome function.

Despite our extensive computational and experimental efforts to assign gene identity to the 2,262 unknown conserved blocks, the evidence we present indicates that most of these conserved blocks are not protein-coding or RNA genes. The selection of ungapped alignments for ≥ 100 bp is conservative, as ungapped sequences are more likely to be coding to maintain the open reading frame. Therefore the number of non-genic conserved sequences we identified here is an underestimate (see also Fig. 1a). Sequence conservation of at least 50% of a sample of the conserved blocks was confirmed in rabbit. This level of conservation, in light of the substitution biases in the sequences described here, shows that the selective constraint is high. If at least 50% of the 2,262 unknown blocks are selectively constrained and 63% of them have no support of being genes, as we have shown, then we estimate that there is a

lower bound of 712 functional sequence blocks of ≥ 100 bp with no gene characteristics on human chromosome 21. This large number of functionally constrained sequences that are not genes probably consists of regulatory regions and unknown functional features of the genome that we are now beginning to discover. Further study of non-genic sequences will improve significantly our understanding of features shared between species as well as species differentiation. Trisomy for some of these non-genic sequences may also contribute to Down's syndrome phenotypes, and their functional characterization provides an attractive challenge for the understanding of genome function in health and disease. Although the present analysis is performed in human chromosome 21, the conclusions can be projected to the whole human and mouse genomes. As discussed in the accompanying manuscript¹² there is a high abundance of conserved regions not corresponding to known genes and a large fraction of those are not genic. The identification of the role of such sequences will probably reveal important clues for genome function and regulation. □

Methods

Alignment

Human sequence was obtained from NCBI (human genome version: build 28) and corresponded to the following contigs: NT_011512 (gi: 161170824), NT_030187 (sequence version gi: 16166615), NT_030188 (gi: 16166749), NT_011515 (gi: 16166537). Mouse genomic sequence was originally obtained from Celera genomics (ref. 7; see also <http://www.celera.com>) through subscription available at the Ludwig Institute for Cancer Research in Lausanne (CVJ). When the publicly available mouse genome was released all 3,491 conserved blocks were subjected to BLAST analysis to verify their presence in the public domain¹². The accession numbers for the mouse sequences syntenic to human chromosome 21 are available as Supplementary Information. Both human and mouse sequences were masked for human and rodent repeats respectively with RepeatMasker (A. F. A. Smit and P. Green, unpublished data (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>)).

Human and mouse sequences were searched with BLAST to identify the genomic boundaries of known genes, and using these boundaries the contigs were separated in slightly overlapping segments of approximately 2–3 Mb. Subsequently they were individually aligned using PipMaker. In all cases where we used BLAST for the identification of the characteristics of the conserved blocks, we required stringent matching criteria with an e -value $< 10^{-20}$ and similarity $> 98\%$.

Gene prediction

We used four different methods to obtain gene models. For the GrailEXP method (<http://compbio.ornl.gov/grailexp/>) we performed exon prediction on extended versions of the human sequence of the conserved blocks by adding 50 bp upstream and downstream to provide the appropriate genomic context, and we chose pairs of predicted exons that were less than 100 kb apart and that had the same orientation as putative exons of a gene. For the Pro-Gen method we carried out homology-based gene prediction using the genomic sequences of human and mouse corresponding to regions with high density of conserved blocks (≥ 5 blocks within 40 kb). For the EST matches pairs of adjacent conserved blocks that matched the same human or mouse EST were considered putative pairs of exons of the same gene. The EST entries used were from the Human Genome Mapping Project (HGMP) database. The fourth method was adjacent conserved blocks. Pairs of adjacent conserved blocks that had a low ratio of replacement (K_a) to silent substitutions (K_s) in one of the six putative coding frames ($K_a/K_s < 0.5$) and similarity between human–mouse $> 85\%$ were chosen.

RT-PCR analysis

The expression of the above predictions was tested by RT-PCR. Total RNA derived from 20 different normal human adult tissues (brain, heart, kidney, spleen, liver, stomach, colon, small intestine, muscle, lung, testis, placenta, skin, peripheral blood lymphocytes, bone marrow, fetal brain, fetal liver, fetal kidney, fetal heart and fetal lung) was extracted, reverse-transcribed and normalized. The quality of total RNA was tested by PCR using MLH1 primers located at intronic sequences flanking exon 12 (forward, 5'-TGGTGTCTCTAGTTCTGG-3'; reverse, 5'-CAITGTTGTAGTAGCTCTGC-3'), as an indicator of possible genomic DNA contamination. Primers for RT-PCR were designed using the Primer3 program (<http://www-genome.wi.mit.edu/cgi-bin/primer/primer3-www.cgi>). Primers were designed from the sequence of distinct putative exons so that the possible amplification of genomic DNA could be distinguished from cDNA amplification. We chose a single PCR rather than a nested PCR approach to avoid false-positive results due to illegitimate transcription. Similar amounts of the 20 cDNAs (final dilution $\times 250$) were mixed with JumpStart REDTaq ReadyMix (Sigma) and 4 ng ml⁻¹ of each of the primers (Sigma-Genosys) with a BioMek 2000 robot (Beckman). The ten first cycles of PCR amplification were performed with a touchdown annealing temperature decreasing from 60 °C to 50 °C; annealing temperature of the next 35 cycles was 50 °C. Amplimers were separated on Ready to Run precast gels (Pharmacia), and positives were sequenced directly.

Microarray data analysis

We identified all of the positive oligonucleotides with the threshold values $R = 13$ and $D = 12Q$ (ref. 21). R and D are threshold values for the ratio and the difference between perfect match intensity and mismatch intensity, respectively. Thus varying these values gives different measures of sensitivity and specificity. We then used BLAST to identify conserved blocks that corresponded to at least two positive oligonucleotides so as to reduce the number of false positives.

Received 16 September; accepted 30 October 2002; doi:10.1038/nature01251.

- Hardison, R. C., Oeltjen, J. & Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7**, 959–966 (1997).
- O'Brien, S. J. *et al.* The promise of comparative genomics in mammals. *Science* **286**, 458–462 (1999) 479–481.
- Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).
- Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
- Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
- Frazer, K. A. *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11**, 1651–1659 (2001).
- Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
- Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Antonarakis, S. E., Lyle, R., Deutsch, S. & Raymond, A. Chromosome 21: a small land of fascinating disorders with unknown pathophysiology. *Int. J. Dev. Biol.* **46**, 89–96 (2002).
- Raymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420**, 582–586 (2002).
- Schwartz, S. *et al.* PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).
- Waterston, R. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- DeSilva, U. *et al.* Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.* **12**, 3–15 (2002).
- Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
- Davisson, M. T. *et al.* Evolutionary breakpoints on human chromosome 21. *Genomics* **78**, 99–106 (2001).
- Gardiner, K., Slavov, D., Bechtel, L. & Davisson, M. Annotation of human chromosome 21 for relevance to down syndrome: gene structure and expression analysis. *Genomics* **79**, 833–843 (2002).
- Raymond, A. *et al.* From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. *Genomics* **78**, 46–54 (2001).
- Raymond, A. *et al.* Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79**, 824–832 (2002).
- Novichkov, P. S., Gelfand, M. S. & Mironov, A. A. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17**, 1011–1018 (2001).
- Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BioMed Central Bioinformatics* **2**, 8 (2001).
- Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
- Nekrutenko, A., Makova, K. D. & Li, W. H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12**, 198–202 (2002).
- Madsen, O. *et al.* Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610–614 (2001).
- Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
- Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).
- Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
- Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
- Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nature Rev. Genet.* **2**, 549–555 (2001).
- Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808 (2002).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements This project was supported by grants from the Swiss National Science Foundation, National Center for Competence in Research 'Frontiers in Genetics', the European Union/Federal office of Education and Health 'Child Care' foundation (to S.E.A.), and a Swiss National Science Foundation grant (to P.B.). We thank E. Lander for advice and support, C. Rossier for core sequencing support and J. Yang for providing programs.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.E.A. (e-mail: stylianos.antonarakis@medecine.unige.ch).

Human chromosome 21 gene expression atlas in the mouse

Alexandre Raymond^{*†}, Valeria Marigo^{†‡}, Murat B. Yaylaoglu^{†§}, Antonio Leoni[‡], Catherine Ucla^{*}, Nathalie Scamuffa^{*}, Cristina Caccioppoli[‡], Emmanouil T. Dermitzakis^{*}, Robert Lyle^{*}, Sandro Banfi[‡], Gregor Eichele[§], Stylianos E. Antonarakis^{*} & Andrea Ballabio^{‡||}

^{*} Division of Medical Genetics, University of Geneva Medical School and University Hospital of Geneva, CMU, 1, rue Michel Servet, 1211 Geneva, Switzerland

[‡] Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, 80131 Naples, Italy

[§] Max Planck Institute of Experimental Endocrinology, Feodor-Lynen-Str. 7, D-30625 Hannover, Germany

^{||} Medical Genetics, Second University of Naples, Naples, Italy

[†] These authors contributed equally to this work

Genome-wide expression analyses have a crucial role in functional genomics. High resolution methods, such as RNA *in situ* hybridization provide an accurate description of the spatiotemporal distribution of transcripts as well as a three-dimensional 'in vivo' gene expression overview^{1–5}. We set out to analyse systematically the expression patterns of genes from an entire chromosome. We chose human chromosome 21 because of the medical relevance of trisomy 21 (Down's syndrome)⁶. Here we show the expression analysis of all identifiable murine orthologues of human chromosome 21 genes (161 out of 178 confirmed human genes) by RNA *in situ* hybridization on whole mounts and tissue sections, and by polymerase chain reaction with reverse transcription on adult tissues. We observed patterned expression in several tissues including those affected in trisomy 21 phenotypes (that is, central nervous system, heart, gastrointestinal tract, and limbs). Furthermore, statistical analysis suggests the presence of some regions of the chromosome with genes showing either lack of expression or, to a lesser extent, co-expression in

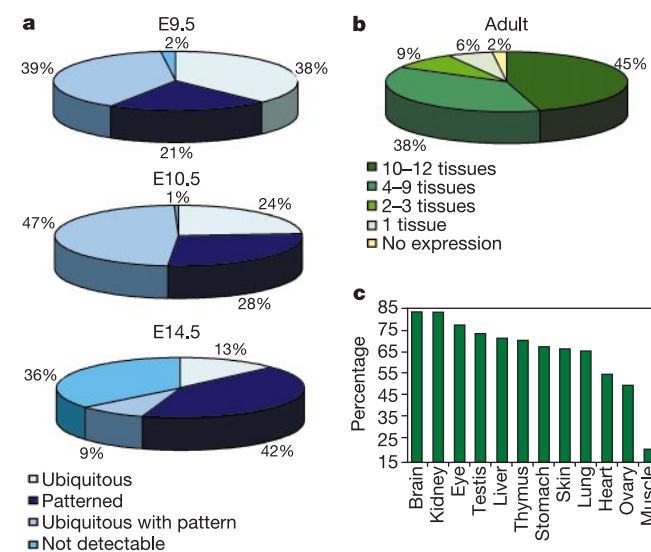


Figure 1 Distribution of expression patterns and transcriptome complexity. **a**, Each slice corresponds to the percentage of genes belonging to the four categories of expression pattern observed by ISH at E9.5 (whole mount), E10.5 (whole mount) and E14.5 (sections). **b**, Each slice represents the percentage of genes expressed in 0, 1, 2–3, 4–9 and 10–12 adult tissues by RT-PCR. **c**, Percentage of the analysed 161 human chromosome 21 murine orthologues identified in each murine adult tissue.

either lack of expression or, to a lesser extent, co-expression in specific tissues. This high resolution expression 'atlas' of an entire human chromosome is an important step towards the understanding of gene function and of the pathogenetic mechanisms in Down's syndrome.

So far 178 confirmed genes and 36 predicted genes have been identified on human chromosome 21 (refs 7–11). The mouse syntenic regions (segments of mouse chromosomes 10, 16 and 17) harbour 170 orthologues. We isolated 237 complementary DNA fragments representing 158 mouse orthologues (93%) for *in situ* hybridization (ISH) experiments and designed primer pairs for 161 genes for polymerase chain reaction with reverse transcription (RT–PCR). To generate the human chromosome 21 gene expression atlas, these orthologues were studied by normalized RT–PCR in 4 developmental stages and 12 adult tissues; whole-mount ISH of embryonic day E9.5 and E10.5 embryos; and ISH on serial sagittal sections of E14.5 embryos (see Supplementary Information for detailed Methods; see also <http://www.tigem.it/ch21exp/>). For ISH of sections, we developed an ISH robot and an automated microscope that permitted the analysis of about 6,500 tissue sections generated for this atlas¹².

Expression was detected for 98% of the tested genes by at least one of the selected methods. The results are compiled in the Supplementary Information and at <http://www.tigem.it/ch21exp/>, and are also summarized in Fig. 1a–c. Supplementary Information consists of three items: (1) atlas expression map tables, containing a list of the genes ordered by their position on chromosome 21 and all original ISH images, annotation tables and details on probes; (2) some examples of patterned expressions in gut, cerebellum, heart, thymus, pancreas and limbs; and (3) a list of Methods. Previously described expression patterns are in agreement with our data (for example, *Sh3bgr*¹³).

By ISH, patterned (regional) gene expression was observed for 21% (E9.5), 28% (E10.5) and 42% (E14.5) of genes (see examples in Fig. 2). The highest numbers of genes with a restricted expression pattern were observed in the brain, the eye and the gut at all stages. Ubiquitous expression was observed in 38% (E9.5), 24% (E10.5) and 13% (E14.5) of cases. Genes with both weak ubiquitous expression and strong regional expression were also observed (39% at E9.5, 47% at E10.5, and 9% at E14.5; for example, *Pfkl*, Fig. 2). No expression was detected for 2% (E9.5), 1% (E10.5) and 36% (E14.5) of genes.

In addition to ISH, we carried out 2,576 RT–PCR reactions covering 95% of the human chromosome 21 murine orthologues (RT–PCR results are documented on Supplementary Information and <http://www.tigem.it/ch21exp/>). The 161 orthologues analysed were found to be expressed on average in 8 of 12 adult tissues tested (s.d. = 3.6). The transcriptomes of brain and kidney showed the highest complexity, each tissue expressing 85% of the 161 genes. Other tissues are less complex (muscle, 21%; heart, 56%; ovary, 51%; lung 67%; skin, 68%; stomach, 69%; thymus, 72%; liver, 73%; testis 75%; and eye, 79%; Fig. 1c). Forty-five per cent of all genes were widely expressed (>9 tissues out of 12 tissues were positive; Fig. 1b). Thirty-eight per cent of the genes show expression in 4–9 tissues. Restricted expression (2–3 tissues) and single-tissue expression was observed in 9% and 6% of the genes, respectively. Only 2% of the genes were not detectable by RT–PCR.

There is good concordance between the whole-mount data at E9.5 and E10.5 (Fig. 1a). It is, however difficult to compare these results with the data obtained by the two other methods, (that is, sections at E14.5 and RT–PCR on adult tissues). This is due to differences in sensitivity (RT–PCR is more sensitive), resolution (sections have a cellular resolution), and ascertaining background signal compared with weak ubiquitous expression (easier in whole

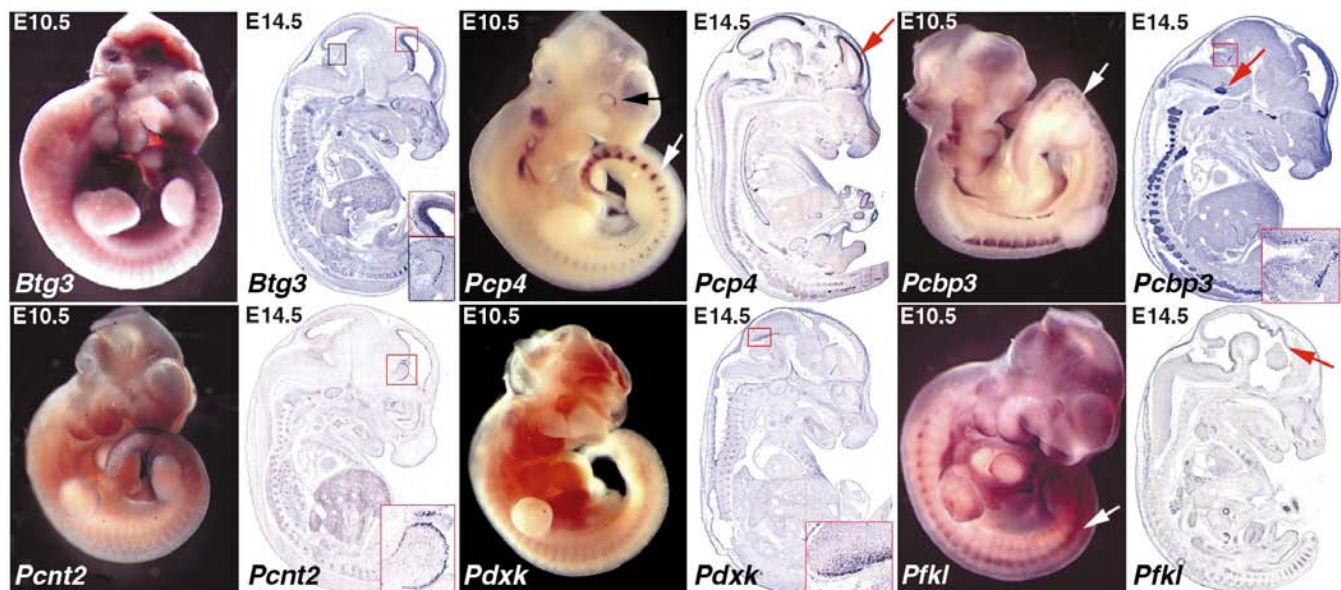


Figure 2 Representative examples of *in situ* hybridization data of E10.5 (whole mount) and E14.5 (sections) embryos. See supplementary Information for the list of genes studied and images of their expression patterns. *Btg3* messenger RNA is ubiquitously expressed at E10.5, with higher abundance in brain and gut. At E14.5 expression in the brain is restricted to the ventricular zone (red insert) and to a group of cells migrating into the developing cerebellum (black insert). *Pcp4* is transcribed in brain, eye (black arrow) and dorsal root ganglia (white arrow) at E10.5. At E14.5 cells transcribing this gene are found in numerous tissues including the cortical plate (red arrow), midbrain, cerebellum, spinal chord, intestine, heart and dorsal root ganglia. *Pcbp3* transcripts are restricted to the central and peripheral nervous systems both at E10.5 and E14.5 (white arrow, dorsal root

ganglia; red arrow, facial nerve nucleus). Insert shows a group of neuroblasts migrating from the midbrain into the cerebellar anlage. Expression of *Pcnt2* at E10.5 is ubiquitous but stronger in brain, eye, limbs, branchial arches and gut. At E14.5 *Pcnt2* brain expression is restricted to proliferating cells of the ventricular zone (red insert). At E10.5 *Pdxk* mRNA is ubiquitous but is more strongly expressed in brain and eye. At E14.5 a group of cells in the inferior colliculus express *Pdxk* (red insert). *Pfkl* transcripts are found throughout the E10.5 embryo with higher levels in brain, eye and dorsal root ganglia (white arrow). At E14.5 this transcript is widely expressed but exhibits regional upregulation in the ventricular zone (red arrow), axial and cranial cartilage, thymus, gut and lung.

mounts). Moreover, E14.5 has a more complex tissue architecture than earlier stages. A longitudinal analysis of 147 of the analysed genes at E9.5, E10.5 and E14.5 revealed that 59% of the genes retain their expression pattern during these developmental stages, whereas 17% acquire new, specific expression at E14.5. For the remaining 24%, we are uncertain owing to low expression in some stages.

Figure 2 presents examples of high-resolution expression patterns in the developing nervous system. Some of the brain-expressed genes may contribute to the Down's syndrome cognitive defects. At E14.5 *Btg3*, *Pcnt2* and *Pfkl* transcripts are detected in the ventricular zone, whereas *Pcp4* staining is observed in the cortical plate and the mantle layer of the midbrain. *Btg3* and *Pcbp3* staining was also observed in neuronal precursors that migrate from the inferior colliculus into the cerebellum. At the same developmental stage, *Pdxk* is expressed in a group of cells in the inferior colliculus. The KH-domain (maxi-K Homology domain) encoding *Pcbp3* transcript is restricted to the central and peripheral nervous systems both at E10.5 and E14.5. Finally, *Pcp4*, *Pcbp3* and *Pfkl* are strongly expressed in dorsal root ganglia, as visible in whole-mount ISH at E10.5.

Down's syndrome is associated with congenital heart disease, and

provides an important model to link individual genes to pathways controlling heart development. In trisomy 21 the most frequent and specific heart abnormalities are atrioventricular canal and atrial septal defects. *Pwp2h* and *C21orf11* show elevated expression in the developing atria (Fig. 3), and *Pfkl* is strongly expressed in the ventricular wall and atrium. Two other heart-specific expression patterns are noteworthy: *Adarb1* and *C21orf18* are both expressed in E10.5 aortic sac; the precursor of the ascending aorta and the pulmonary artery. Furthermore, *C21orf18* is expressed in the bulbus cordis, which develops into the right ventricle. At E14.5, *Kcnj15* and *Adarb1* are expressed in the aortic valve and trunk, while *Kcnj15* is also expressed along the outflow track of the heart and in the superior vena cava. *Atp50* and *Sh3bgr* transcripts are detected throughout the heart, whereas *Cldn8* is regionally expressed in the primitive ventricle. The human genes *COL6A1*, *COL6A2*, *COL18A1* and *KCNE2* are mutated in Bethlem myopathy, Ullrich's disease, long QT6, and Knobloch syndrome, respectively^{14–18}. We found that *Kcne2* is expressed in the entire developing heart including its vessels, whereas *Col18a1* is detected in cardiac vessels only. *Col6a1* and *Col6a2* are strongly expressed in the mitral valve and along the pericardium^{16,18} (Fig. 3).

Down's syndrome fetuses exhibit reduced growth rate of long

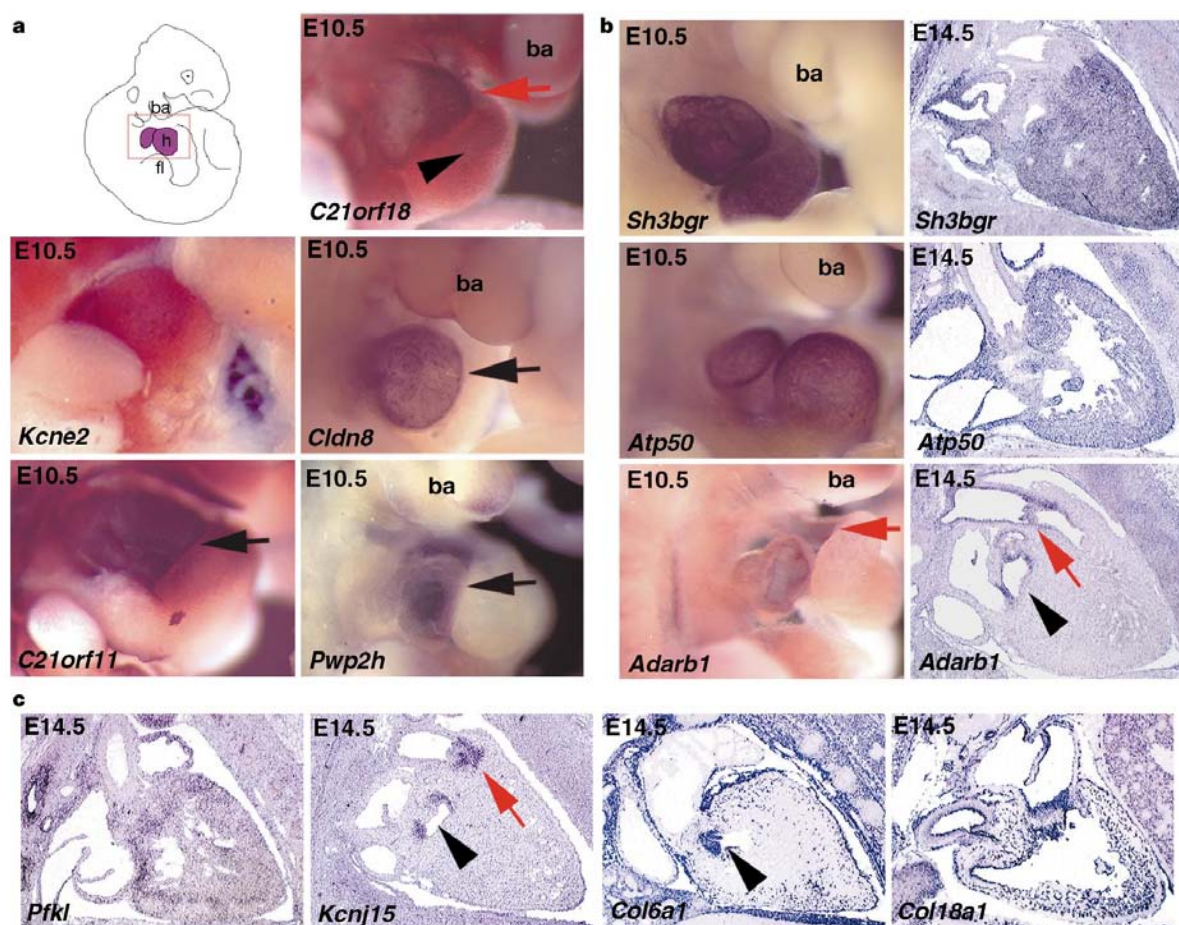


Figure 3 Expression analysis in the developing heart. E10.5 whole embryos and E14.5 sections were probed for *Adarb1*, *Atp50*, *C21orf11*, *C21orf18*, *Col6a1*, *Col18a1*, *Cldn8*, *Kcnj15*, *Kcne2*, *Pfkl*, *Pwp2h* and *Sh3bgr* genes. **a**, The region portrayed from the whole-mount embryos is schematically represented in the red box of the top left panel. h, heart; ba, branchial arch; fl, forelimb. At E10.5, *C21orf18* transcripts are present in the aortic sac (red arrow) and the bulbus cordis (black arrowhead). *Kcne2* is expressed in atria and ventricles. mRNA for *C21orf11* and *Pwp2h* is restricted to the atria (black arrows), and *Cldn8* transcripts are restricted to the primitive ventricle (black arrow). **b**, *Sh3bgr* and

Atp50 transcripts are detected throughout the heart. *Adarb1* mRNA is present in the aortic sac (red arrow) at E10.5, and is restricted to the mitral valve (black arrowhead), aortic valve (red arrow) and the endothelium of the aortic trunk at E14.5. **c**, At E14.5 *Pfkl* is expressed throughout the heart. *Kcnj15* transcripts are found in the aortic and mitral valve (red arrow and black arrowhead, respectively). *Col6a1* is strongly expressed in the mitral valve (black arrowhead) and along the pericardium, whereas *Col18a1* expression is seen in the vessels of the heart.

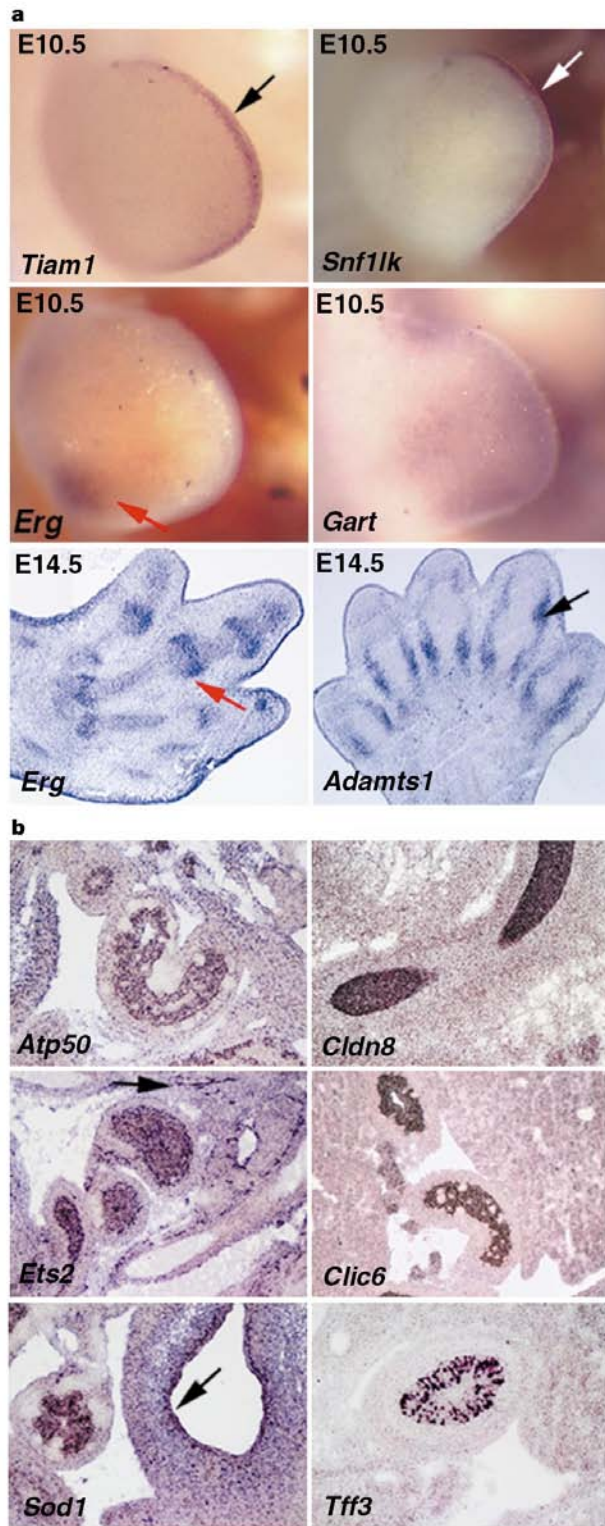


Figure 4 Expression analysis in the developing limb and gastrointestinal tract. **a**, At E10.5 *Tiam1* and *Snf1lk* are expressed specifically in the apical ectodermal ridge, which is a specialized ectodermal thickening regulating proliferation of the underlying mesoderm (arrows). A more diffuse distribution in the limb bud mesoderm was found for *Gart*. *Erg* transcripts are present in the posterior proximal mesoderm at E10.5 and in the joints (red arrows) at E14.5. *Adamts1* showed expression in the perichondrium of the developing digits (black arrow). **b**, *Atp50*, *Cldn8*, *Ets2* and *Clic6* are expressed within the duodenum endoderm. *Ets2* transcripts are also detected in the vasculature surrounding the duodenum (arrow). *Sod1* mRNA is present in the midgut endoderm and in the stomach endothelium (arrow). *Tff3* is transcribed by a subgroup of cells within the gastroduodenal junction region of the pyloric sphincter.

limb bones during the third trimester of pregnancy. Furthermore, a non-ossified or hypoplastic middle phalanx of the fifth digit is present in these fetuses⁶. Numerous genes are expressed in the limb buds (Fig. 4a). *Tiam1* and *Snf1lk* transcripts were found specifically in the apical ectodermal ridge; a specialized ectodermal thickening regulating proliferation of the underlying mesoderm (Fig. 4a). *Erg* showed strong expression in the posterior proximal mesoderm at E10.5 and in joints at E14.5. Finally, *Adamts1* was expressed at E14.5 in the perichondrium of developing bones.

Gastrointestinal abnormalities such as duodenal stenosis, Hirschsprung's disease, gastro-oesophageal reflux and imperforate anus are frequent in Down's syndrome patients⁶. At E14.5 the digestive tract is well differentiated and the expression of many genes can be detected. *Atp50*, *Cldn8*, *Clic6* and *Ets2* are expressed within the endothelium of the duodenum (Fig. 4b). Interestingly, the latter is also expressed in the skeletal system and previously found to result in skeletal abnormalities reminiscent of Down's syndrome when overexpressed in transgenic mice¹⁹. *Tff3* and *Sod1* transcripts are present in a subgroup of cells within the gastroduodenal junction region of the pyloric sphincter, and in the endothelium of the stomach and the intraperitoneal portion of the midgut, respectively.

Previous studies suggested the presence of genomic regions containing clusters of genes with similar expression patterns^{20–24}. To search for such clusters on chromosome 21, all RT-PCR expression data underwent a test for clustering²⁵ (see Methods in Supplementary Information). In four regions we found significant clustering of genes showing absence of expression in specific tissues. These include genes not expressed in heart (from *B3gal5* to *Hsf2bp*; 3.9 Megabases (Mb), 31 genes, $P < 0.001$), lung (from *Dscam* to *Slc37a1*; 2.7 Mb, 19 genes, $P = 0.007$), testis (from *C21orf25* to *Pde9a*; 1.0 Mb, 12 genes, $P = 0.028$), and muscle (from *Hsf2bp* to *Hrmt1l1*; 3.4 Mb, 43 genes, $P = 0.02$). The following duplicated genes were found: *Mx1* and *Mx2*; *Tff1*, *Tff2* and *Tff3*; *Ifnar1* and *Ifnar2*; and *Col6a1* and *Col6a2*. Notably, each of these regions is fully contained within a syntenic block on the mouse chromosomes. Consistent clustering results were obtained when the cluster analysis was extended to expression data collected by ISH. In addition, E14.5 ISH data suggested a cluster for presence of expression within fore-, mid- and hindbrain, which extended from *Hunk* to *Atp50* (22 genes, $P = 0.004$). Significant clustering was observed even when applying two additional statistical analyses: the Bonferoni correction and Fisher's combined probability (see Supplementary Information). Albeit preliminary, these data suggest that regions containing either co-silenced or co-expressed genes exist on chromosome 21, and their presence should be considered in future expression studies of large genomic regions.

The combination of gene mapping with expression analysis is a helpful tool in the identification of positional candidate genes for human diseases²⁶. Thus the human chromosome 21 expression atlas provides a rich resource for candidate genes for both monogenic and multifactorial diseases mapping to chromosome 21. We anticipate, however, that the greatest impact of the atlas lies in assessing the contribution of specific genes to Down's syndrome traits and phenotypes. The correlation of the atlas expression patterns with specific features of Down's syndrome may lead to the identification of candidate genes for these features of the disease. In particular, *ADARB1*, *KCNJ15*, *PFKL*, *PWP2H* and *SH3BGR* are promising candidate genes for Down's syndrome heart defects, as they map to the Down's syndrome congenital heart disease critical region²⁷, and because of the expression patterns of their murine orthologues in the developing heart. Similarly, the human orthologues of the gut-expressed genes *Atp50*, *Cldn8*, *Clic6*, *Ets2*, *Hmg1*, *Sh3bgr*, *Sod1* and *Wrb* may have a role in the Down's syndrome gastrointestinal abnormalities⁶. Mapping gene expression of an entire chromosome at high resolution defines a new level of gene annotation, which is anticipated to advance our knowledge on gene function and

regulation, and our understanding of human aneuploidies, such as Down's syndrome. □

Received 11 May; accepted 19 September 2002; doi:10.1038/nature01178.

- Strachan, T., Abitbol, M., Davidson, D. & Beckmann, J. S. A new dimension for the human genome project: towards comprehensive expression maps. *Nature Genet.* **16**, 126–132 (1997).
- Neidhardt, L. *et al.* Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos. *Mech. Dev.* **98**, 77–94 (2000).
- Bulfone, A. *et al.* The embryonic expression pattern of 40 murine cDNAs homologous to *Drosophila* mutant genes (Dres): a comparative and topographic approach to predict gene function. *Hum. Mol. Genet.* **7**, 1997–2006 (1998).
- Reymond, A. *et al.* The tripartite motif family identifies cell compartments. *EMBO J.* **20**, 2140–2151 (2001).
- Gawantka, V. *et al.* Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev.* **77**, 95–141 (1998).
- Epstein, C. J. *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 749–794 (McGraw Hill, New York, 1995).
- Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Davison, M. T. *et al.* Evolutionary breakpoints on human chromosome 21. *Genomics* **78**, 99–106 (2001).
- Pletcher, M. T., Wiltshire, T., Cabin, D. E., Villanueva, M. & Reeves, R. H. Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics* **74**, 45–54 (2001).
- Reymond, A. *et al.* Nineteen additional unpredicted transcripts from the human chromosome 21. *Genomics* **79**, 824–832 (2002).
- Reymond, A. *et al.* From PREDs and open reading frames to cDNA isolation: revisiting the human chromosome 21 transcription map. *Genomics* **78**, 46–54 (2001).
- Herzig, U. *et al.* Development of high-throughput tools to unravel the complexity of gene expression patterns in the mammalian brain. *Novartis Found. Symp.* **239**, 129–146 (2001).
- Egeo, A. *et al.* Developmental expression of the SH3BGR gene, mapping to the Down syndrome heart critical region. *Mech. Dev.* **90**, 313–316 (2000).
- Abbott, G. W. *et al.* MiRP1 forms IKr potassium channels with HERG and is associated with cardiac arrhythmia. *Cell* **97**, 175–187 (1999).
- Sertie, A. L. *et al.* Collagen XVIII containing an endogenous inhibitor of angiogenesis and tumour growth, plays a critical role in the maintenance of retinal structure and in neural tube closure (Knobloch syndrome). *Hum. Mol. Genet.* **9**, 2051–2058 (2000).
- Camacho Vanegas, O. *et al.* Ullrich scleroatonic muscular dystrophy is caused by recessive mutations in collagen type VI. *Proc. Natl. Acad. Sci. USA* **98**, 7516–7521 (2001).
- Higuchi, I. *et al.* Frameshift mutation in the collagen VI gene causes Ulrich's disease. *Ann. Neurol.* **50**, 261–265 (2001).
- Jobis, G. J. *et al.* Type VI collagen mutations in Bethlem myopathy, an autosomal dominant myopathy with contractures. *Nature Genet.* **14**, 113–115 (1996).
- Sumarsono, S. H. *et al.* Down's syndrome-like skeletal abnormalities in Ets2 transgenic mice. *Nature* **379**, 534–537 (1996).
- Kim, S. K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
- Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* **6**, 6 (2002).
- Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26**, 183–186 (2000).
- Bortoluzzi, S. *et al.* A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* **8**, 817–825 (1998).
- Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1**, 5 (2002).
- Tang, H. & Lewontin, R. C. Locating regions of differential variability in DNA and protein sequences. *Genetics* **153**, 485–495 (1999).
- Ballabio, A. The rise and fall of positional cloning? *Nature Genet.* **3**, 277–279 (1993).
- Barlow, G. M. *et al.* Down syndrome congenital heart disease: a narrowed region and a candidate gene. *Genet. Med.* **3**, 91–101 (2001).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank M. Traditi and G. Lago for the design of the website, and B. Fischer for preparation of the specimens. We are grateful to F. Chapot, S. Deutsch, M. Guipponi, K. Hashimoto, P. Kahlem, J. Michaud, H. S. Scott and M. L. Yaspo for plasmids and reagents, and to J. Aidan, M. Friedli, C. Rossier and the Telethon Institute of Genetics and Medicine (TIGEM) RNA *in situ* hybridization core for core assistance. This work was supported by grants from the Jérôme Lejeune Foundation to R.L. and A.R.; from the Swiss Fonds National Suisse de la Recherche Scientifique, the European Union/Office Fédéral de l'Éducation et de la Santé and ChildCare foundation to S.E.A.; from the German Ministry of Research to G.E.; from the EC Fifth Framework Program to A.B. and G.E.; from the Italian Telethon Foundation to TIGEM; from The National Center for Competence in Research-Frontiers in Genetics to S.E.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B. (e-mail: ballabio@tigem.it), S.E.A. (e-mail: stylianos.antonarakis@medecine.unige.ch) or G.E. (e-mail: gregor.eichele@mpih.mpg.de).

Authors' contributions The three laboratories of A.B., S.E.A. and G.E. contributed equally to this work.

A gene expression map of human chromosome 21 orthologues in the mouse

The HSA21 expression map initiative

***Group 1:**Yorick Gitten†, Nadia Dahmane‡, Sonya Baik†

Ariel Ruiz i Altaba†

***Group 2:**Lorenz Neidhardt§, Manuela Scholze§, Bernhard G. Herrmann§

***Group 3:**Pascal Kahlem||, Alia Benkahl||, Sabine Schrinner||, Reha Yildirimman||, Ralf Herwig||, Hans Lehrach|| & Marie-Laure Yaspo||

* These author groups contributed equally to the work.

† Skirball Institute, Developmental Genetics Program and Department of Cell Biology, New York University School of Medicine, 540 First Avenue, New York, New York 10016, USA

‡ Université de la Méditerranée – CNRS UMR 6156, Institut de Biologie du Développement, Campus de Luminy, Case 907, 13 288 Marseille Cedex 09, France

§ Max Planck Institute for Immunology, Department of Developmental Biology, Stübweg 51, D-79108 Freiburg, Germany

|| Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics, Ihnestrasse 73, D-14195 Berlin, Germany

The DNA sequence of human chromosome 21 (HSA21)¹ has opened the route for a systematic molecular characterization of all of its genes. Trisomy 21 is associated with Down's syndrome, the most common genetic cause of mental retardation in humans. The phenotype includes various organ dysmorphies, stereotypic craniofacial anomalies and brain malformations². Molecular analysis of congenital aneuploidies poses a particular challenge because the aneuploid region contains many protein-coding genes whose function is unknown. One essential step towards understanding their function is to analyse mRNA expression patterns at key stages of organism development. Seminal works in flies, frogs and mice showed that genes whose expression is restricted spatially and/or temporally are often linked with specific ontogenic processes. Here we describe expression profiles of mouse orthologues to HSA21 genes by a combination of large-scale mRNA *in situ* hybridization at critical stages of embryonic and brain development and *in silico* (computed) mining of expressed sequence tags. This chromosome-scale expression annotation associates many of the genes tested with a potential biological role and suggests candidates for the pathogenesis of Down's syndrome.

The current HSA21 gene catalogue¹ (<http://chr21.molgen.mpg.de>) contains 238 entries for which we have identified 168 cognate mouse orthologues (see Methods in the Supplementary Information). We isolated 187 mouse cDNA clones matching 158 unique genes (referred to as mmu21 genes; see Supplementary Table 1) whose orthology was confirmed on the basis of the known synteny between HSA21 and segments of mouse chromosomes MMU16, MMU10 and MMU17 (<http://www.informatics.jax.org/>) (see Supplementary Information; this database is also available at <http://chr21.molgen.mpg.de/hsa21/>).

To identify potential candidates with a role in patterning and organ development, we have explored the expression of the 158 mmu21 genes by systematic whole-mount *in situ* hybridization³ (WISH) at mid-gestation (embryonic day 9.5; E9.5), a stage covering a wide range of embryonic processes. We also analysed a subset of clones at two other stages (Supplementary Table 1). We found 111 of 158 genes expressed at E9.5; 78 genes showed widespread expression and 33 genes a restricted pattern (49% and 21%, respectively, of the genes examined at E9.5). In addition, 12 widespread genes also defined particular embryonic structures (for example, *Prkcbp2*; Supplementary Table 1). Among the mmu21 genes conserved in *Saccharomyces cerevisiae* (Y), *Caenorhabditis*

elegans (C) and *Drosophila melanogaster* (D) (CDY in Supplementary Table 1), 60% were widely expressed and 17% showed a restricted pattern. In contrast, 43% of the genes conserved only in multicellular organisms (CD) were widespread and 30% were patterned. The functional classification of the 45 pattern genes shows a bias for cell–cell communication and signal transduction molecules. Only six of the mmu21 genes were previously analysed by whole mounts^{4–9}, including *Ets2*, a gene that causes skeletal defects in transgenic mice¹⁰. Our analysis at E9.5 identified a number of interesting patterns (Fig. 1). *Kiaa0184* is a new gene with a striking expression profile in the central and peripheral nervous systems, marking subsets of cells in the midbrain and hindbrain, basal plate margin cells of the spinal cord, and cranial as well as dorsal root ganglia. *Samsn1*, containing SH3 and SAM domains, was found associated with blood vessel formation. In addition, we uncovered new potential roles for previously characterized genes. *Tsga2*, a component of junctional membrane complexes reported to be testis-specific¹¹, is restricted to rhombomeres at E9.5 and might take part in early events of hindbrain patterning. Informative patterns might be associated with organ-specific traits possibly found in Down's syndrome, although their function remains to be demonstrated. *Igsf5* and *Tff3* are candidates for male sterility. *Igsf5* is present in the mesonephros required for seminiferous tubule formation. *Tff3*, a secretory protein promoting cell migration in intestinal epithelium^{12,13}, was detected in a small group of cells possibly representing primordial germ cells. *Sh3bgr* is strikingly expressed in heart, correlating with the specific cardiac defects in Down's syndrome. Genes (such as *Slc19a1*, *Clic6* and *Lss*) expressed in the

branchial apparatus might have a role in the stereotypic facial abnormalities. *Adamts5*, a disintegrin metalloprotease, is specifically expressed in the maxillary process of the branchial apparatus, and transiently (E8.75–9.5) in the ventral midbrain, making it a candidate for facial abnormalities and mental retardation in Down's syndrome. Of the 28 genes expressed in the brain and/or head mesenchyme at E9.5, 20 show a regionalized expression in the maturing brain at postnatal day 2 (P2) (Supplementary Table 1); examples are *Clic6* and *Tsga2*, which are restricted to the choroid plexus (Fig. 2).

We have investigated the expression of mmu21 genes in neonatal brain, the most important organ affected in Down's syndrome. At P2, the maturation of brain structures involves critical events, including a major shift from neurogenesis to gliogenesis and the elaboration of cytoarchitecture and connectivity. Under our experimental conditions and scoring criteria, 60% of the mmu21 genes were found expressed by *in situ* hybridization (ISH) in the P2 mouse brain (Supplementary Table 1). Of the clones expressed in brain, 42% gave widespread signals, half of these also displaying regionally enhanced expression (for example, *Pfkl*; Supplementary Table 1); 58% of the expressed clones defined only patterns in discrete brain territories, as illustrated in Fig. 2. Figure 3 summarizes the distribution of the mmu21 transcriptome in 11 mouse brain regions.

Overall, there was a prominent expression of mmu21 genes in post-mitotic cells (Fig. 2), whereas only 19 genes were found expressed in the mitotically active ventricular/subventricular zones (Supplementary Table 1). The largest fraction of the mmu21 transcripts expressed at P2 was detected in three major laminar structures of the brain: neocortex (41%), hippocampus (25%) and cerebellum (25%). These structures develop through partly common mechanisms^{14,15} and share the expression of a

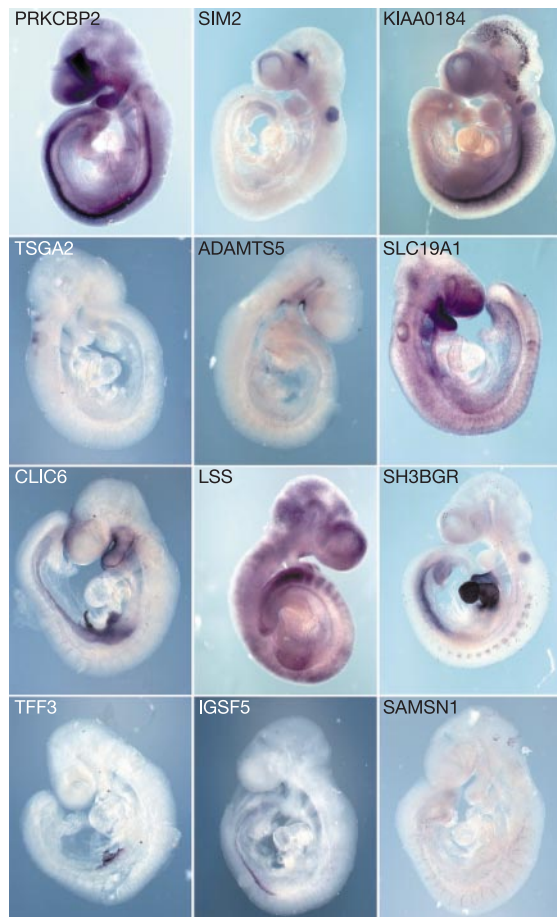


Figure 1 Examples of expression patterns of mmu21 genes in whole-mount E9.5 mouse embryos. Gene symbols of human orthologues are used; for details see Supplementary Table 1.

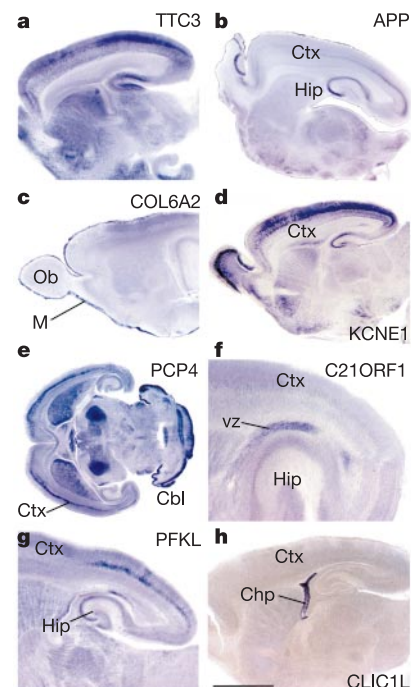


Figure 2 Expression pattern of selected genes on postnatal brain sections by RNA *in situ* hybridization. All sections are from P2 brains and are sagittal, except **e**, which is horizontal. Specific expression is denoted in blue–purple. Anterior is to the left and dorsal side is up. Hip, hippocampus; Ctx, cortex; Tct, tectum; Cb, cerebellum; Ob, olfactory bulb; Chp, choroid plexus; M, meninges; VZ/SVZ, ventricular zone/subventricular zones. Scale bar in **h** represents 200 μ m in **a–e**, **g**, **h** and 70 μ m in **f**. Gene symbols of human orthologues are used.

significant number of genes (Fig. 3). We observed a bias towards neocortical expression, although it remains unclear whether such predominance is also found for the genes encoded by other chromosomes^{16,17}. Transcripts prominently expressed in this structure represent a pool of candidates for the cognitive defects of Down's syndrome. *Dscam*, *Synj1* and *Tiam1* have already been described as being involved in synaptic function, axonal guidance, cell migration and neurite outgrowth^{18–20}. The hippocampus is necessary for the integrity of memory, and impaired spatial memory in Down's syndrome²¹ might be associated with some of the hippocampal genes, such as the semaphorin-like oncogene *Pttg1p* (ref. 22). A plot of the brain expression domains along the chromosome (Fig. 3a) did not show evidence for the existence of local expression clusters. We found four genes expressed in the meninges (*Wrb*, *Col18a1*, *Col6a1* and *Col6a2*), a tissue enveloping the central nervous system. The three collagen genes are physically close on 21q22.3 and might represent a functionally regulated gene cluster. Because many of the mmu21 genes are expressed in the dorsal brain at P2, our data would be compatible with the hypothesis that cumulative gene dosage effects might underlie brain malfunction in trisomy 21 (refs 23–25).

We obtained a wider overview of mmu21 gene expression by extracting *in silico* expression data from sequenced mouse cDNA libraries representing a large range of tissues and developmental stages. Screening of 632 libraries from non-cancerous tissues totaling 2,253,129 mouse expressed sequence tags (ESTs) identified

5,454 ESTs (on average 34 EST hits per mmu21 gene) displayed in Supplementary Table 1 (see Methods in the Supplementary Information), allowing a detailed analysis of mmu21 gene expression. Although EST mining shows limitations for measuring differential gene expression levels^{26–29}, we can exploit EST counts to reveal patterns of genes with correlated expression profiles by using a correlation-based clustering method³⁰. Figure 4 shows the expression profiles for the mmu21 genes, many of which are weakly or moderately expressed in more than one tissue. Interestingly, 19 genes seemed prominently expressed in single tissues (red boxes in Fig. 4), suggesting that they might have a crucial role in a tissue-specific mechanism (for example, *Tff* (refs 12, 13)). Gene clusters identified here provide information on possible roles of these genes in common biological mechanisms. For instance, five genes (*Runx1*, *Aire*, *Ifngr2*, *Itgb2* and *Ubash3a*) from cluster no. 8 showing thymus-specific expression are linked to immunity or haematopoiesis.

We sought to compare ISH data with expression profiles generated by other approaches. Genes that were scored as not expressed (or not interpretable) on brain sections were tested by polymerase chain reaction with reverse transcription (RT-PCR) on whole P2 brain. As expected, RT-PCR was more sensitive: almost 50% of these ISH-negative genes were found expressed in whole P2 brain by this method (Supplementary Table 1). It is likely that most of those are either expressed ubiquitously at very low levels or are present within a few cells that were absent from the tested sections, which

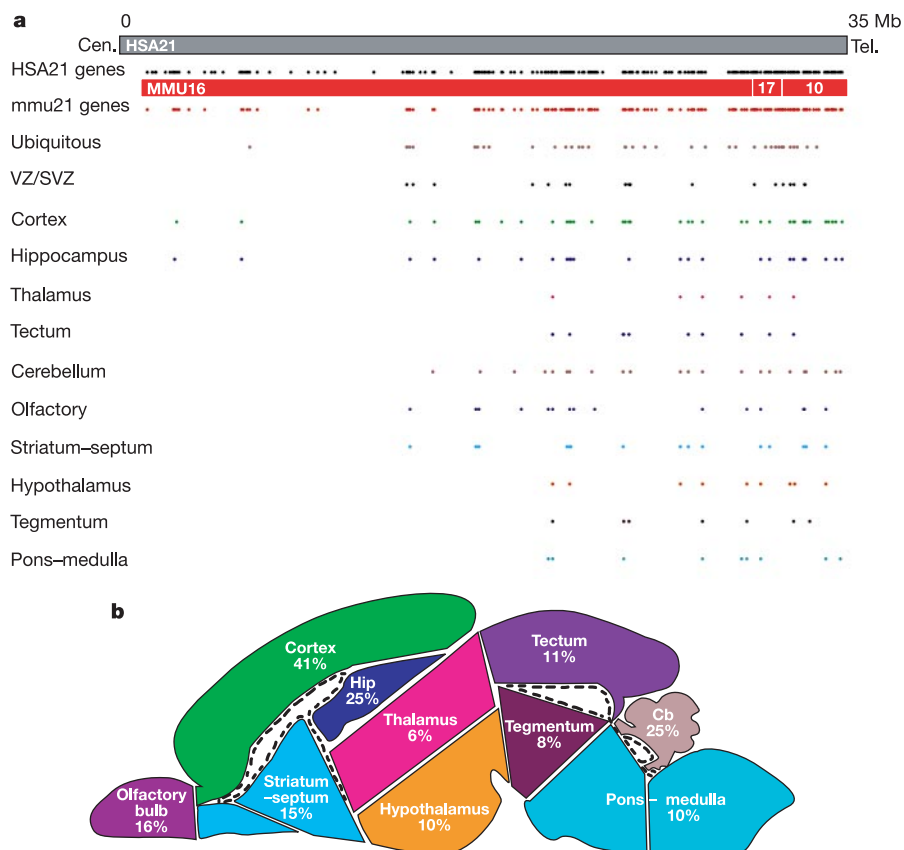


Figure 3 Expression of mmu21 genes in P2 mouse brain. **a**, Chromosomal distribution of the HSA21 genes (top row, black), of their mouse orthologues (second row, red) and of the expression domains of mmu21 genes in different brain regions. The reference scale is that of HSA21 genomic sequence; horizontal bars represent the mouse syntenic chromosomal segments (MMU16, MMU17 and MMU10). Cen, centromere; Tel, telomere; VZ/SVZ, ventricular/subventricular zones. **b**, Diagram summarizing the expression of

mmu21 genes in different regions of the P2 mouse brain determined by ISH as reported in **a**. The brain is shown as a sagittal section divided into its major developmental and functional domains. Percentages in each domain refer to the proportion of clones showing expression in each area out of the expressed clones. The ventricular zone is shown by black dashed line. Cb, cerebellum; Hip, hippocampus.

cannot cover the whole brain even though median sagittal sections span a maximum of structures. ISH and RT-PCR are complementary, because ISH aims at identifying patterned genes that are often highly expressed in a given territory, whereas RT-PCR lacks topographical resolution. In addition, the efficiency of riboprobes depends in part on their sequence composition or on the occurrence of splice variants that might show differential expression. It was difficult to correlate our expression map of the P2 brain with the EST information, owing to the heterogeneity in size, tissue region and developmental stages of the original material for the different brain EST libraries analysed. Of the genes found expressed by section ISH, 60% did identify brain ESTs (interrogating 251,426 brain ESTs pooled for all brain regions and ages ranging from neonatal to adult, excluding embryo; see Supplementary Information); 65% of the ISH⁺/PCR⁺ genes hit brain ESTs, which originated from stages older than P2 with a few exceptions, for example *Ifnrg2* and *Cryz11* at neonatal stage. Of the genes scoring ISH⁺/PCR⁺ at P2, 37% hit brain ESTs originating mostly from adult stages, except *Cbs*, which is present in the neonatal cortex. Finally, 16 genes are not, or are very weakly, expressed in neonatal brain because they scored negative by all three methods (*Cldn8*, *Kcne1*, *Runx1*, *Sim2*, *Wdr9*, *C21orf11*, *Mx2*, *Mx1*, *C21orf25*, *Tff3*, *Tff2*, *Ubash3a*, *Wdr4*, *Kiaa0958*, *Ftcd* and *Hsf2bp*).

As expected, the correlation between data from WISH and EST

mining was very good because both techniques are able to detect transcripts in most or all embryonic tissues (620,471 ESTs ranging from stage 2 cells to E19.5). This correlation was quantified statistically with Fisher's exact test to refute the hypothesis of statistical independence of the two techniques. For each gene we denoted whether it had positive or negative expression in both methods. Out of 158 genes tested, 123 showed an agreement either with expression (100) or lack of expression (23) in both techniques, whereas in 35 cases we found a disagreement ($P = 2.0496 \times 10^{-7}$). Of the 47 genes not detected at E9.5 by WISH, 23 are unlikely to be expressed at early stages of development because they do not hit any embryonic EST either. One-quarter are found only expressed in libraries originating from later stages, starting from E12.5–E14.5. Ten genes hit ESTs encompassing stage E9.5 (*Tmprss2*, *Slc37a1*, *Cbs*, *Tmem1*, *Dnmt3l*, *C21orf19*, *Ifnar2*, *Cldn8*, *Usp16* and *Usp25*) but were not detected, possibly because of very low expression levels in E9.5 embryos. Overall, most of the transcripts detected at E9.5 by WISH did hit ESTs and also showed sustained expression until stages E14.5–E15.5 or later. Nevertheless, embryonic EST libraries are often derived from the whole fetus and therefore lack organ information. Where information is available, EST mining is correlated with the WISH results (for example, *Sh3bgr* in heart, or *Kiaa0184* and *Pcbp3* in brain).

We observed local clusters of gene expression in several chromo-

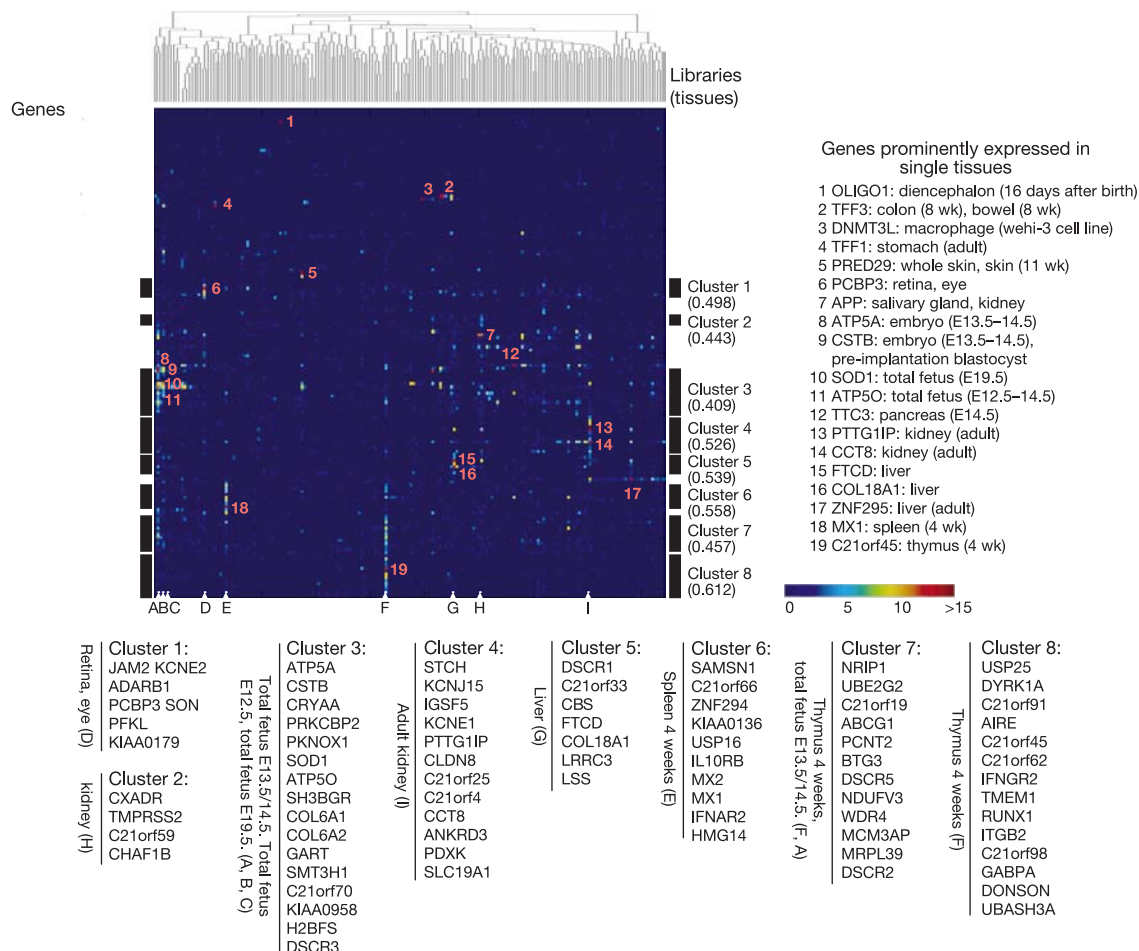


Figure 4 EST analysis. Matrix displaying the expression profiles of 159 mmu21 genes (rows) within 190 cDNA libraries (columns). Dendrograms used to reorder libraries (top) and genes (left) are shown, together with eight significant gene clusters (solid bars, left and right of the matrix together with correlation coefficients). The cluster composition is shown with corresponding libraries (A–I) and indicated by white arrows at the bottom of

the matrix). Coloured dots represent the number of ESTs found in a given library for each gene (see scale). Numbers in red refer to genes prominently expressed in single tissues, listed at the right. Interactive figure and details are given in the Supplementary Information. Gene symbols of human orthologues are used.

somal regions. Significance was judged by computing upper-tail *P* values (<0.01 in all cases) for observed numbers of expressed genes in groups of neighbouring genes of given sizes in accordance with a hypergeometric distribution. WISH data suggest co-expression from *Pdxk* to *Mcm3ap* in head mesenchyme and extra-embryonic component, and from *Tmprs2* to *Abcg1* in the nose. EST mining data give evidence for a significant group from *H2bfs* to *Pfkl* co-expressed in haematopoietic tissues, kidney, liver and mammary gland. In addition, a large silenced region from *Kcne2* to *Cryaa* was identified in brain, ovary, retina and testis.

Our analysis of the pattern of gene expression for HSA21 is based on complementary techniques, combining the thorough coverage of developmental stages by EST mining with the high anatomical resolution of WISH and brain ISH. We have identified new expression patterns for a large number of genes, including many with highly localized expression, and generated information on the possible functions for a large fraction of the HSA21 genes. This study provides tools and insights for the further analysis of developmental processes and identifies a pool of gene candidates for phenotypes specific to Down's syndrome and other pathologies linked to HSA21. □

Received 11 June; accepted 12 September 2002; doi:10.1038/nature01270.

1. The chromosome 21 mapping and sequencing consortium The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
2. Epstein, C. J., et al. *The Metabolic and Molecular Bases of Inherited Disease*, 7th edn (ed. Scriver, C. R.) 749–794 (McGraw-Hill, New York, 1995).
3. Neidhardt, L. et al. Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos. *Mech. Dev.* **98**, 77–94 (2000).
4. Maroulakou, I. G., Papas, T. S. & Green, J. E. Differential expression of *ets-1* and *ets-2* proto-oncogenes during murine embryogenesis. *Oncogene* **9**, 1551–1565 (1994).
5. Lin, B. et al. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease *TMPRSS2*. *Cancer Res.* **59**, 4180–4184 (1999).
6. Egeo, A. et al. Developmental expression of the *SH3BGR* gene, mapping to the Down syndrome heart critical region. *Mech. Dev.* **90**, 313–316 (2000).
7. Epstein, D. J. et al. Members of the bHLH-PAS family regulate Shh transcription in forebrain regions of the mouse CNS. *Development* **127**, 4701–4709 (2000).
8. Orti, R. et al. Characterization of a novel gene, *C21orf6*, mapping to a critical region of chromosome 21q22.1 involved in the monosomy 21 phenotype and of its murine ortholog, *orf5*. *Genomics* **64**, 203–210 (2000).
9. Casas, C. et al. *Dscr1*, a novel endogenous inhibitor of calcineurin signaling, is expressed in the primitive ventricle of the heart and during neurogenesis. *Mech. Dev.* **101**, 289–292 (2001).
10. Sumarsono, S. H. et al. Down's syndrome-like skeletal abnormalities in *Ets2* transgenic mice. *Nature* **379**, 534–537 (1996).
11. Taketo, M. M. et al. Mapping of eight testis-specific genes to mouse chromosomes. *Genomics* **46**, 138–142 (1997).
12. Mashimo, H., Wu, D. C., Podolsky, D. K. & Fishman, M. C. Impaired defense of intestinal mucosa in mice lacking intestinal trefoil factor. *Science* **274**, 262–265 (1996).
13. Otto, W. R. & Patel, K. Trefoil factor family (TFF)-domain peptides in the mouse: embryonic

- gastrointestinal expression and wounding response. *Anat. Embryol.* **199**, 499–508 (1999).
14. Altman, J. & Bayer, S. A. *Atlas of Prenatal Rat Brain Development* (CRC Press, Boca Raton, 1995).
15. Dahmane, N. et al. The Sonic Hedgehog–Gli pathway regulates dorsal brain growth and tumorigenesis. *Development* **128**, 5201–5212 (2001).
16. Sandberg, R. et al. Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl Acad. Sci. USA* **97**, 11038–11043 (2000).
17. Miki, R. et al. Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl Acad. Sci. USA* **98**, 2199–2204 (2001).
18. Leeuwen, F. N. et al. The guanine nucleotide exchange factor Tiam1 affects neuronal morphology: opposing roles for the small GTPases Rac and Rho. *J. Cell Biol.* **139**, 797–807 (1997).
19. Yamakawa, K. et al. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum. Mol. Genet.* **7**, 227–237 (1998).
20. Saito, T. et al. Mutation analysis of *SYNJ1*: a possible candidate gene for chromosome 21q22-linked bipolar disorder. *Mol. Psychiat.* **6**, 387–395 (2001).
21. Sylvester, P. E. The hippocampus in Down's syndrome. *J. Ment. Defic. Res.* **27**, 227–236 (1983).
22. Pei, L. Identification of *c-myc* as a down-stream target for pituitary tumour-transforming gene. *J. Biol. Chem.* **276**, 8484–8491 (2001).
23. Delabar, J. M. et al. Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.* **1**, 114–124 (1993).
24. Korenberg, J. R. et al. Down syndrome phenotypes: the consequences of chromosomal imbalance. *Proc. Natl Acad. Sci. USA* **91**, 4997–5001 (1994).
25. Kahlem, P. & Yaspo, M. L. Human chromosome 21 sequence: impact for the molecular genetics of Down syndrome. *Gene Funct. Dis.* **1**, 175–183 (2000).
26. Schmitt, A. O. et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27**, 4251–4260 (1999).
27. Vingron, M. & Hoheisel, J. Computational aspects of expression data. *J. Mol. Med.* **77**, 3–7 (1999).
28. Scheurle, D. et al. Cancer gene discovery using digital differential display. *Cancer Res.* **60**, 4037–4043 (2000).
29. Krause, A., Haas, S. A., Coward, E. & Vingron, M. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.* **30**, 299–300 (2002).
30. Ewing, R. M. et al. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9**, 950–959 (1999).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Sultan, H.-J. Warnatz, G. Teltow, B. Eppens, D. Balzereit, C. Bail, V. Palma, P. Sánchez, M. Beyna, J. Mullor and M. Rallu for their contribution; S. Antonarakis for providing clones; P. Avner for the mouse somatic cell hybrids; and the Resource Center of the German Genome Project (RZPD, Berlin) for providing clones and filters. Y.G., N.D., L.N., P.K. and A.B.K. contributed equally. This work was supported by grants from the Hirscht Trust and the NIH-NINDS and NCI to A.R.A.; from an ATIPE grant to N.D.; from the German Human Genome Project to B.G.H.; from the National Genome Forschung Netz (NGFN), the Ministry of Education and Research (BMBF) and the Max Planck Society to M.-L.Y.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.R.A. (e-mail: ria@saturn.med.nyu.edu), B.G.H. (e-mail: herrmann@immunbio.mpg.de) or M.-L.Y. (e-mail: yaspo@molgen.mpg.de).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Hideki Watanabe

e-mail: h.watanabe@naturejpn.com

naturejobs

An electoral lecture

The classic whispered advice to new professors goes something like this: focus on researching and publishing; spending too much time on teaching can hamper your career. "I've heard that," says Dennis Jacobs, professor of chemistry at the University of Notre Dame in Indiana. But he's chosen to ignore it. Jacobs was last month named US professor of the year for research and doctoral universities by the Council for Advancement and Support of Education and the Carnegie Foundation for the Advancement of Teaching.

Over the years, Jacobs has tweaked his courses to make them less reliant on lectures and memorizing facts. For example, in his introductory chemistry course he talks briefly about a concept — how various bonds are formed, for example. Before he performs an experiment to demonstrate the concept, he asks students to predict an outcome — perhaps whether the temperature of the solution will rise, fall or stay the same. He then asks them to discuss their theory with their neighbours and vote. Next, he canvasses the lecture hall for different answers. After the opinions are aired, he asks for another vote, performs the experiment and discusses the outcome.

Preparing this type of lecture takes a lot of work, and his colleagues have asked whether it is worth the effort. So, after overhauling his course for the first time a few years ago, he tried his new approach on only one group, and compared long-term outcomes of the students with the modified and standard teaching styles. Not surprisingly, students that he has tracked from the modified class did better overall in other chemistry classes than those who received the more standard instruction.

Perhaps such data — and awards that acknowledge similar successes — will also turn teaching into something that is emphasized, rather than swept aside.

Paul Smaglik
Naturejobs editor



Contents

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS